Journal of Biomedical Informatics 93 (2019) 103159

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A comprehensive review of feature based methods for drug target interaction prediction

Kanica Sachdev*, Manoj Kumar Gupta

Computer Science and Engineering Department, SMVDU, J&K, India

ARTICLE INFO

Keywords:
Drugs
Drug target interaction
Feature based techniques
Proteins
Targets

ABSTRACT

Drug target interaction is a prominent research area in the field of drug discovery. It refers to the recognition of interactions between chemical compounds and the protein targets in the human body. Wet lab experiments to identify these interactions are expensive as well as time consuming. The computational methods of interaction prediction help limit the search space for these experiments. These computational methods can be divided into ligand based approaches, docking approaches and chemogenomic approaches. In this review, we aim to describe the various feature based chemogenomic methods for drug target interaction prediction. It provides a comprehensive overview of the various techniques, datasets, tools and metrics. The feature based methods have been categorized, explained and compared. A novel framework for drug target interaction prediction has also been proposed that aims to improve the performance of existing methods. To the best of our knowledge, this is the first comprehensive review focusing only on feature based methods of drug target interaction.

1. Introduction

Drug target interaction refers to the binding of a drug to a target location that results in a change in its behavior/function. A drug or medicine basically refers to any chemical compound which brings about a physiological change in the human body when it is consumed, injected or absorbed. Target, also known as biological target, is any part of the living organisms to which the drugs bind in order to bring the physiological change. Targets are the entities like proteins or nucleic acids which are directed for any change. The most common biological targets are nuclear receptors, ion channels, G-protein coupled receptors and enzymes. Drug target interaction prediction plays a vital role in the drug discovery process which aims to identify new drug compounds for biological targets. Fig. 1 shows the process of drug target interaction. As shown in the figure, the chemical compound of the drug binds to the target molecule by forming temporary bonds. The attached drug then reacts with the biological target to create a positive or a negative change and consequently leaves the biological target. The drugs inhibit the functioning of the target to prevent certain catalyzed reactions occurring in the human body in order to treat diseases. This is achieved by inhibiting its contact with certain enzymes called substrates. The drug target interaction can occur in two ways. The first kind of drugs, known as competitive inhibitors, attach themselves to the active site of the target to impede the reaction. The second type of drugs, called allosteric inhibitors bind to the allosteric site of the target. This changes the shape and structure of the target which averts the substrate from recognizing it. Thus, the reactions do not occur. The blocking of the target's reactions can correct metabolic imbalance or kill pathogens to cure diseases.

The drug target interactions can be inferred by wet lab experiments using various techniques of classical and reverse pharmacology [4]. However, laboratory experiments conducted to predict these interactions are time consuming as well as costly. Thus, in-silico prediction of interactions between drugs and target proteins is greatly desirable [6]. The computational techniques can effectively predict the possible interactions with great accuracy, thereby reducing the search space to be investigated in laboratory experiments.

There are various factors that have made the drug target interaction prediction more necessary in the current scenario. Over the past decade, a large number of compounds have been synthesized. However, their target profiles and drug effects are still unidentified. Moreover, there are a number of diseases like Lichen planus [8], Parkinson's disease [11] etc., which have no cure and many new diseases are being introduced every year. Also, researchers have accumulated massive information on various compound properties, features, responses and target proteins to construct numerous large scale datasets. As such, researchers need to develop more efficient methods to manipulate and analyze these high dimensional, complex data. Thus, there is an urgent

* Corresponding author.
E-mail addresses: kanica.sachdev@gmail.com (K. Sachdev), manoj.gupta@smvdu.ac.in (M.K. Gupta).

https://doi.org/10.1016/j.jbi.2019.103159
Received 13 November 2018; Received in revised form 25 March 2019; Accepted 26 March 2019
Available online 27 March 2019
1532-0464/

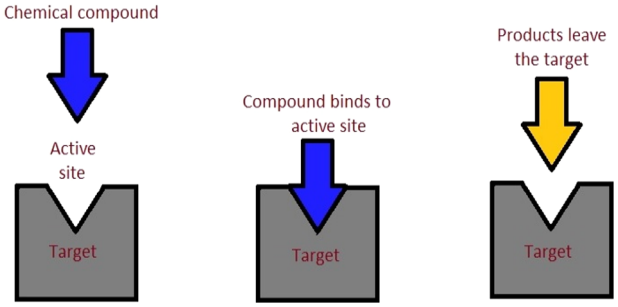

Fig. 1. The process of drug target interaction.

need to develop more accurate and powerful computational methods for prediction of drug target interaction [13].

Drug target interaction prediction has various applications. It facilitates the process of drug discovery [15], drug repositioning [16–18] and drug side effect prediction [21,22]. Drug discovery is the exploration of new drugs that interact with a particular target. In-silico drug target interaction prediction can help identify new drugs that bind to these targets. Discovering a new drug is expensive as well as an inefficient technique. It involves various stages like identifying a specific target that binds to a chemical compound, determining the lead compound of the chemical that binds to the target protein and lead optimization to enhance the efficiency as well as the specificity [24]. Following all these steps, the drugs then undergo various clinical trials before being introduced in the market. The cost of identifying each new molecular entity (NME) is approximately $1.8 billion [26]. Moreover, newly developed drug compounds almost take about a decade to reach the consumable market. For example, the US Food and Drug Administration approves only about 20 new drugs in a year [28]. As such, the drug discovery mechanism is tedious and prolonged. In addition, a huge number of already known chemical compounds are not currently being used as drugs as their interaction profiles with proteins are not known yet. The PubChem dataset, for instance, contains the details of more than 90 million compounds, a majority of which do not have recognized interaction profiles. Detecting these interaction profiles by in-silico methods would aid the identification of new drugs and would also help to narrow down the drug search space to work within the process of drug development.

Additionally, some drugs are shown to interact with multiple target sites, known as polypharmacology. The earlier drug discovery followed the principle of “One molecule- one target- one disease”, which implied discovering specific drug molecules that bind to specific targets for treating specific diseases. Under this paradigm, a particular protein is selected and studied for a certain therapeutic action. This has helped in identifying effective chemical compounds that act on particular proteins. Although this aids in designing drugs that act on specific factors of a disease, they do not provide a solution for complex diseases that are multifactorial as well as vulnerable at multiple attacks. Recent discoveries have shown that in certain cases, removing a certain gene individually causes no effect. However, removing two genes forming a pair leads to harmful effects. This implies that such pairs of genes compensate for each other’s effects. Since multiple genes are responsible for treating a particular disease; the drugs should affect each of these genes to be effective against their related diseases. Such diseases demand the manipulation of multiple factors simultaneously along multiple stages in order to be treated effectively. Also, multiple targets may be involved in a particular disease. Single target drugs may not produce the desired effect on the entire biological system as their effectiveness may be inhibited in compensatory ways [30]. Thus, the detection of multi-target drugs as well as drug combination research helps in the treatment of various complex diseases. This has also led to the development of drug repositioning which aims at using the known drugs or compounds to treat a different disease. The study of known drugs and their interaction profiles can discover multiple benefits of a single drug. These drugs can then be used to treat other diseases as well. For example, Gleevec (imatinib mesylate) was initially thought to be interactive with the Bcr-Abl fusion gene that was related to the disease leukemia. Later, it was also found to interact with PGDF and KIT. Gleevec was then repositioned to heal gastrointestinal stromal tumors also [33].

Polypharmacology also suggests that a drug which may have a therapeutic effect on a target may produce a side effect on some other target simultaneously. Drug side effects are the adverse effects that a drug may produce in addition to treating some other problem. They have become a serious problem in the domain of drug discovery. As an example, drug side effects are assumed to cause approximately 100,000 deaths per year in the US causing it to be the fourth largest cause of deaths [35]. These side effects are the major cause of drug failure. Early prediction of drug side effects at clinical stages will help to improve the laborious and expensive process of drug testing and will also provide safe and efficient therapies for patients [22]. All these applications highlight the role of drug target interaction in the process of drug development and research. Thus computational approaches for the prediction of drug target interactions can tremendously aid drug development.

Polypharmacology has aided the process of drug repositioning or drug repurposing, which implies the use of an already discovered drug for treating other diseases. This process has several advantages. Firstly, an already approved drug has undergone extensive clinical trials before being introduced in the market. Thus, it is already established that the drug is safe for use. Also, the clinically approved drugs have been analyzed in a comprehensive manner. Thus, their various properties, therapeutic effects, side effects etc are already known. This accelerates the study of these drugs and can help identify their interactions efficiently [14]. Thus, repurposing drugs can help save time and effort immensely.

There are various factors that determine which in-silico method should be employed for interaction prediction. The most important factor is the stage of drug development [24]. In the initial stages, for instance, the focus of the study would be on the disease to be analyzed. It would involve the identification of disease genes or the differentiation of healthy and infected genes. The later stages can involve processes like selecting a lead compound for the chemical that can help in optimizing the disease treatment. Subsequently, Quantitative Structure-Activity Relationships (QSAR) studies are conducted to determine the drug properties of the lead compound [36]. The method selected also depends on the type of data that is available. The information accessible for the drug and target properties will greatly determine the method to be employed [37].

There are also various problems related to drugs that adversely affect the process of drug discovery and interaction prediction. A single drug may have multiple effects. It is very difficult to identify and enumerate all of these effects, which include positive as well as negative consequences. Also, a drug may have different responses even though the genes may be almost similar [38,39]. Moreover, the biological pathways are very complex and complicated. Hence, it is very difficult to identify relevant interactions from them [27].

The computational methods for predicting drug target interactions can be broadly classified into three categories: ligand based approaches, docking approaches and chemogenomic approaches [6]. Ligand based approaches are developed on the idea that similar molecules usually bind to similar protein targets and show similar properties [40]. Thus, these approaches predict the interactions based on the similarities between the protein ligands. One such method like QSAR uses machine learning methods to compare a candidate ligand with the set of all known ligands to infer its binding capability [41,42]. This approach, however, suffers from certain disadvantages. Since the sequence information of the proteins is not used for prediction, the possible novel interactions are limited to the link between known ligands and protein

families. Also, the performance of these methods decreases considerably when the known ligands for a particular candidate protein are less [3].

The second category i.e. docking approach uses the 3D structures of the proteins as well as drugs to run a simulation in order to predict if they would interact [43–45]. This approach is also prone to some drawbacks. There are certain proteins like membrane proteins whose 3D structures are not known since the prediction of their structures is a challenging task. Docking approaches cannot be applied in these cases [46]. Also, when receptor proteins are considered, the method becomes very complex as the number of degrees of freedom to be considered is very large [47].

The third category known as chemogenomic approaches, utilize the information of drugs and proteins simultaneously in order to make interaction predictions. The chemical space of drugs and the genomic space of proteins are unified in a common subspace to infer possible interactions. The chemogenomic approaches overcome the disadvantages of ligand based and docking approaches that have been discussed previously. The basic advantage of this approach is that it can use extensive biological data that is readily available in public datasets. For instance, the chemical structure graphs and genomic sequences have already been compiled in public online databases. These graphs and sequences are used in possible interaction prediction [23].

The chemogenomic methods can further be broadly classified into two categories: feature based methods and similarity based methods [48]. The feature based methods represent the drug target pair with a vector of descriptors. The various properties of drugs, as well as the proteins, are encoded as corresponding features. The feature based methods predict the interactions between these drug target pairs by discovering features that are more discriminative. Thus, the inputs of these methods are various feature vectors that may be produced by combining the properties of drug and targets. These feature vectors are then fed into various machine learning models like random forest, Support Vector Machines (SVM) etc to predict novel interactions.

On the other hand, a variety of similarity based methods have also been developed to calculate the similarity between the drug compounds and target proteins. Various kernel functions have been defined that make use of similarity matrices. These matrices are computed using different techniques and measures.

The rest of the paper is organized as follows. Section 2 highlights the various tools and terms related to drug target interaction. It mentions the important datasets, tools for calculating descriptors and the numerous metrics that have been proposed for evaluation. Section 3 exhaustively explains the different feature based methods. The methods have been divided into three categories, namely SVM based methods, ensemble based methods and miscellaneous techniques. The techniques falling under each of these three categories have been discussed. Section 4 summarizes the features based techniques and provides a comparison of the methods. Section 5 provides the advantages as well as the limitations of the various techniques while providing certain enhancement propositions. Section 6 proposes a novel framework for drug target interaction prediction. Section 7 concludes the paper while providing future prospects.

2. Datasets, tools and metrics for feature based methods

2.1. Datasets

The drug target interaction is not just a simple binary on-off relationship. It is affected by various factors, such as the concentrations of the two compounds and their intermolecular interactions. The various datasets provide different information regarding these drug compounds, the target proteins and their interactions. This information can be used to predict various novel potential interactions effectively and accurately. Various eminent databases have been enumerated as under:

(1) *DrugBank* [49]: This database is richly annotated, freely available bioinformatics resource that combines the drug information with detailed drug target data. The most recent release of DrugBank contains 11,123 drugs that include 2558 small molecule drugs, 963 biotech drugs, 112 nutraceuticals and more than 5130 experimental drugs. It also contains 5117 protein sequences that are related to these drugs. Each drug contains more than 20 fields of information which contain the chemical data as well as the target/protein data. DrugBank is extensively being used by chemists, pharmacists, researchers as well as the industry and general public.

(2) *KEGG* [50]: Kyoto Encyclopedia of Genes and Genomes (KEGG) is a collection of various genomes and biological pathways. In addition, it also contains information about various diseases, drugs and chemical compounds. It has been further divided into various sub-databases like KEGG GENE, KEGG PATHWAY, KEGG DRUG etc.

(3) *PubChem*[51]: This database is a collection of various chemical compounds and their related activities. It contains data of 93.9 million chemical compounds 236 million substances and 1.25 million Bioassays. The database can be searched and analyzed for various properties like molecular weight chemical formula etc using online editor.

(4) *UniProt* [52]: This database has been specifically developed for proteins. It is an open source resource that contains information about protein sequences and their biological functional information. It is further subdivided into four parts: UniProtKB (consisting of Swiss-Prot and TrEMBL), UniParc and UniRef. Swiss-Prot consists of protein sequences that have been manually annotated. TrEMBL has automatic annotation. UniParc is a non redundant database of protein sequences which assigns a unique identifier to each sequence. UniRef consists of clusters of protein sequences from UniProtKB and UniParc.

(5) *Pfam* [53]: Pfam database has been developed for the protein families. It includes the proteins, their annotations and their sequence alignments. It basically classifies the various proteins under families and domains. These sequence alignments are developed using the hidden Markov models. For each of the protein families, the users can view their family description, their structures, alignments etc.

(6) *SuperTarget* [54]: SuperTarget is a web based dataset that combines the drug related data with side effects, medical indications, drug metabolism, pathways and Gene Ontology for target data. Currently, this dataset contains more than 6000 target proteins, 19,600 drug compounds and 330,000 interactions between them. The web based interface provides options for drug screening as well as target similarity inclusion. The query interface also allows the user to frame complex queries to find particular drugs or targets.

(7) *MATADOR* [55]: This resource contains the chemical-target interactions. It is different from other datasets like the DrugBank as it includes direct as well as indirect interactions. DrugBank, on the other hand, contains only direct interactions. The annotated interactions (direct and indirect) have been integrated using an automated text mining tool. The indirect interactions are caused by several factors like changes in gene expression fall or binding of a metabolite to a drug. All these factors have been integrated in order to capture maximum interactions.

(8) *GLIDA* [56]: GLIDA is a dataset that has been developed specifically for a class of proteins known as G-protein coupled receptors (GPCR). Since GPCRs are mainly responsible for controlling various physiochemical changes, they have been explored in this database. It contains the biological information of the GPCRs as well as the chemical information of its various ligands.

(9) *Therapeutic target database (TTD)* [57]: Therapeutic target database is a dataset that contains all the information related to the

known therapeutic proteins and nucleic acid targets, their pathway information, the related diseases and the concerned drugs for each of these targets. It also provides data related to protein 3D structure, sequence, ligand binding properties, enzyme nomenclature, therapeutic class, drug effects and clinical development status. The dataset currently contains data about 433 targets, 809 related drugs for these targets and 125 diseases. The data can be accessed by searching according to drug name, target name, disease name or drug function.

(10) *STITCH* [58]: STITCH refers to Search Tool for Interactions of chemicals. It combines the data about drug target interactions, crystal structures, binding experiments and metabolic pathways. The data that has been integrated from text mining, phenotypic effects and chemical structure similarity of compounds is used to infer the relations between chemicals. It contains data of more than 68,000 chemicals which include 2200 drugs and relates them to 1.5 million genes and their interactions.

(11) *ChEMBL* [59]: ChEMBL is a manually created dataset that contains the details of bioactive molecules with drug-like properties. The database provides bioactivity data against drug targets. The information can be analyzed and studied for lead identification during drug discovery. The latest release of the database contains more than 2.97 million bioassay measurements covering 636,269 compounds.

(12) *TDR Targets* [60]: This database is an open resource that provides the availability of various genomic and chemical datasets in order to help in the identification as well as the prioritization of drugs and targets in disease pathogens. Users can design specific queries in order to prioritize data. TDR Targets contains data for 11 bacterial and eukaryotic pathogens, and more than 800,000 bioactive compounds.

(13) *PDTD* [61]: PDTD is an online database for target identification. It provides information on over 1100 proteins with 3D structures. The information has been integrated from various literature sources as well as the web-accessible databases like TTD and DrugBank. It contains over 830 known protein targets, related diseases and biochemical functions. It provides functions like keyword search using ID, drug name or disease name. The results can also be analyzed using visualization tools.

(14) *SIDER* [62]: Side Effect Resource (SIDER) integrates data on drugs, targets and the side effects of drugs to provide a more comprehensive view of actions of drugs and their adverse reactions. The latest version of the dataset contains information on 1430 drugs, 588 side effects and 140,064 drug-side effect pairs. It also provides a dataset of drug indications.

(15) *Integrity* [63]: Integrity is a repository designed for drug discovery. It contains information about a large number of drugs and their associated targets, diseases and clinical phases. Drugs targets are also assigned a status that may be 'Validated' (launched drugs or under active development in clinical phases), 'Candidate' (that are no longer under active development for a specific disease), 'Exploratory' (under biological exploration) or none. Drugs in Integrity are linked to genes through target IDs.

(16) *NIST Mass Spectral Library* [64]: This is a collection of various peer reviewed databases that have been compiled by the National Institute for Standards and Technology (NIST). It contains the mass spectrometry data of the various chemical compounds.

To allow an easier comparison many authors have used common datasets that were initially introduced by Yamanishi et al. [23]. The drug target links have been considered for four targets namely enzymes (E), ion channels (ICs), nuclear receptor (NR) and GPCR targets from different public datasets. These public datasets include KEGG, BRITE, BRENDA, SuperTarget and DrugBank. This has helped in the easier cross comparison of different machine learning models.

2.2. Tools

There are various online web servers and software packages and toolkits to calculate various chemical descriptors for drugs and properties of different proteins. Some of the packages are discussed as under:

(1) *ChemCPP* [65]: This is a C++ basic tool for calculating the kernel functions between the chemical compounds. The various kernel functions include marginalized graph kernels, Tanimoto kernels, tree based graph kernels etc.

(2) *EDragon* [66]: This is a multiplatform software that has been developed to calculate the molecular descriptors of various chemical compounds. These descriptors can be used in evaluating molecular structure activity, similarity analysis etc. It computes more than 1600 topological and geometrical descriptors.

(3) *CDK Descriptor* [67]: This open source platform can detect and group the chemicals based on descriptor classes. It can also compute various fingerprints like hashed, MACCS etc.

(4) *Open Babel* [68]: This tool is used to convert the chemical files from one format to another. It provides several features like substructure search, calculation of fingerprints etc.

(5) *RDKit* [69]: It is a toolkit for generating various descriptors for chemical compounds. In addition, it also provides features like 2D depiction, molecular serialization, fingerprint generation, similarity analysis etc.

(6) *PaDEL* [70]: This is a software that is used to calculate the chemical descriptors and fingerprints. It computes 1875 descriptors and 12 different types of fingerprints.

(7) *BlueDesc* [71]: BlueDesc is a java tool that is used to compute molecular descriptors. It calculates several topological, geometrical, constitutional and whim descriptors. It also identifies various functional groups.

(8) *ChemDes* [72]: This is a web platform that basically integrates various packages like PaDEL, RDKit, BlueDesc etc. It provides functionalities like format conversion, descriptor calculation, fingerprint generation, similarity calculation etc.

(9) *Rcpi* [73]: Compound-Protein Interaction with R (Rcpi) package is used for complex representations of drugs, proteins and their interactions. It computes various chemical, physiochemical and structural descriptors.

(10) *PyDPI* [74]: Drug-Protein Interaction with Python (PyDPI) is a python package developed specifically for drug discovery. It computes molecular descriptors for drugs and structural and physiochemical properties for proteins.

(11) *protr* [75]: This is a package developed in R in order to represent the various protein representations in numerical format. The various descriptors used in the representation are amino acid composition, Position Specific Scoring Matrix (PSSM), C-Terminal Domain (CTD) etc. It also computes the similarity of protein sequences.

(12) *SPiCE* [76]: This is a web based tool for computing various sequence based features of the proteins. Along with computing the protein features, it also provides tools for data visualization and protein classification.

(13) *propy* [77]: Protein in Python (propy) is a package developed in python to compute the physiochemical and structural properties of the proteins. It also calculates 13 different features including tripeptide composition, sequence-order-coupling number etc and Pseudo Amino Acid Composition (PseAAC) descriptors.

(14) *ProtDCal* [78]: This software developed in Java is used to generate thousands of protein features based on 3D structure and sequence. It also provides features like sequence encoding similarity computation and prediction of function.

(15) *ProtParam* [79]: It is used for calculating various physical and chemical properties of the proteins that are available in SwissProt or TrEMBL database. Various parameters like amino acid

Table 1
Tools for calculating drug protein descriptors.

S. no.	Package	For	Link
1.	ChemCPP	Drugs	chemcpp.sourceforge.net/
2.	EDragon	Drugs	www.vcclab.org/lab/edragon/
3.	CDK Descriptor	Drugs	www.rguha.net/code/java/cdkdesc.html
4.	Open Babel	Drugs	http://openbabel.org/
5.	RDKit	Drugs	http://www.rdkit.org/
6.	PaDEL	Drugs	www.yapcwsoft.com/dd/padeldescriptor/
7.	BlueDesc	Drugs	https://omictools.com/bluedesc-tool
8.	ChemDes	Drugs	www.scbdd.com/chemdes/
9.	Rcpi	Drugs and proteins	http://bioconductor.org/packages/release/bioc/html/Rcpi.html
10.	PyDPI	Drugs and proteins	https://sourceforge.net/projects/pydpicao/
11.	protr	Proteins	http://protr.org
12.	SpiCE	Proteins	http://helix.ewi.tudelft.nl/spice.
13.	Propy	Proteins	http://code.google.com/p/protpy/downloads/list
14.	ProtDCal	Proteins	http://bioinf.sce.carleton.ca/ProtDCal/
15.	ProtParam	Proteins	https://web.expasy.org/protparam/

composition, molecular weight etc are computed.

The various packages and their web links have been compiled in Table 1.

2.3. Metrics

The metrics provide a means of evaluation of various techniques. They help to compare different methods in order to select the optimal method for execution. The various metrics defined for drug target interaction prediction to evaluate and compare various techniques have been listed below:

(1) *Positives and negatives* [2]: Four metrics, namely true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) are used to provide an overview of performance. These can be defined as:
TP is the interacting drug target pair that is predicted to be interacting whereas FP is the non interacting drug target pair that is predicted to be interacting. Similarly, TN is the non interacting drug target pair that is predicted to be non interacting and FN is the non interacting drug target pair that is predicted to be interacting.

(2) *Precision and accuracy* [1]: These are the most basic metrics for the evaluation of various techniques. Accuracy is the fraction of correctly identified interactions of the predictor and precision depicts the proportion of correct positive interactions. The accuracy and precision can be computed as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

(3) *Recall* [5]: Recall shows the proportion of positive interactions which were identified correctly. It can be calculated as:

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

(4) *Sensitivity* [1]: Another common evaluation metric sensitivity can be defined as:

$$Sensitivity = \frac{TP}{TP + FN} \quad (4)$$

(5) *Mathew's Correlation coefficient* [1]: Its value ranges from −1 to 1, where −1 represents a completely wrong binary classifier and 1 represents a perfectly correct binary classifier. Mathew's Correlation coefficient can be defined as:

$$Mathew's\ correlation\ coefficient\ (MCC) = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \quad (5)$$

(6) *Area under curve* [9]: The Receiver Operating Characteristic (ROC) curve depicts the performance of a predictor at various threshold values. The true positive rate vs. false positive rate values are plotted to form the curve. To compare these curves, Area Under the Curve (AUC) is computed. It represents an aggregate of the values at various points on the curve. The value of AUC ranges from 0 to 1. A similar metric, Area Under the Precision-Recall (AUPR) graph is also used for performance evaluation.

(7) *Time* [19]: The time taken for the various classifiers for training as well as prediction purposes can also serve as an evaluation and comparison metric for various techniques.

(8) *Memory usage* [10]: Memory is another important consideration while selecting an efficient algorithm for practical usage. The memory requirement of the different algorithms can be considered for evaluation of techniques.

3. Feature based techniques

The feature based methods calculate feature descriptors for both drugs and targets which are then used for prediction. The drug feature vector $[d_1, d_2, \ldots d_n]$ and the target feature vector $[f_1, f_2, \ldots f_m]$ are calculated independently. These may be computed by selecting certain discriminative properties for encoding or using certain software packages or tools that can compute the descriptors automatically. Some methods also employ dimensionality reduction techniques to reduce feature dimensionality, thereby improving the performance as well as the efficiency of the technique. The features are then combined by simply concatenating the features or using methods like tensor product calculation. The drug target pair vector is then used in a classifier.

For the purpose of classification, the entire dataset is divided into training and testing data. The data consists of two types of drug target interaction pairs: positive and negative. The positive pair implies the drug target pair that is known to interact. The negative pairs are those for which the interaction is not known. The drug target pair vectors of the training data are then transferred to a classifier. Various classifiers like Support Vector Machines (SVM), Relevance Vector Machines (RVM), Rotation Forest (RF) etc have been used under different techniques. These techniques have been discussed in detail in the following sub sections. The classifier then predicts the label for the testing data i.e. whether the drug target pair interacts or not. This process has been diagrammatically explained in Fig. 2.

The feature based techniques can be classified into three main categories: SVM based methods, Ensemble based methods and miscellaneous techniques on the basis of the prediction model employed. These three categories of techniques have been analyzed in Sections 3.1, 3.2 and 3.3. A flowchart depicting the categorization of the chemogenomic techniques into similarity based methods and feature based methods have been shown in Fig. 3. It also shows the further categorization of feature based methods.

3.1. SVM based methods

SVMs have been used for classification and regression analysis in various data mining applications. They classify the data into categories based on the training samples to ease the process of data analysis. SVMs provide a means to classify the drug target pairs into interacting and

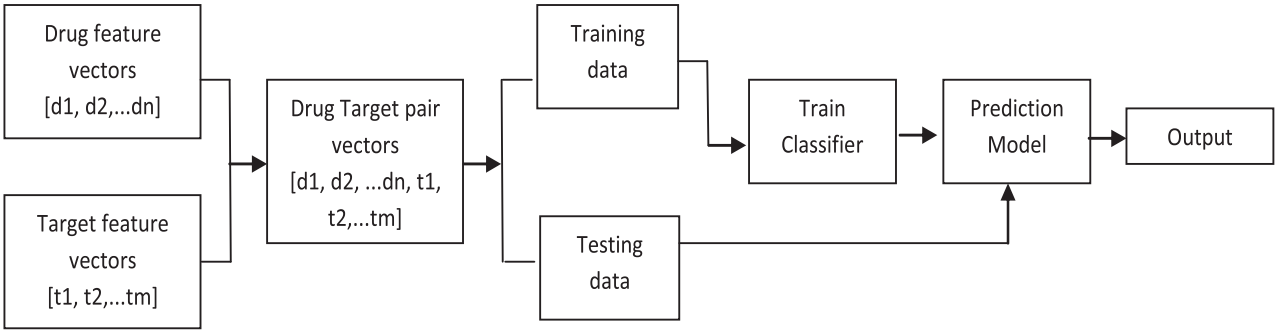

Fig. 2. Overview of steps of feature based methods.

non interacting categories. SVM based methods compute the drug and target features individually. These features then help in the process of classification using a kernel function. The kernel function used for both drugs and targets may be similar or different. SVM based methods have low complexity and are easier to implement. The different SVM based techniques have been discussed as under:

3.1.1. Statistical prediction of protein–chemical interactions based on chemical structure and mass spectrometry data

Previously, drug target interaction was primarily predicted by considering the pharmacological effects of the chemicals. This method introduced drug target interaction prediction using the chemical structure and mass spectrometry of the compounds [1]. For proteins, amino acid sequences are considered. These are used to classify the proteins as well as the chemicals into binding and non binding pairs. The classification is performed using SVM to achieve accurate performance on large quantities of data. The relevance of SVM in the field of prediction and statistical analysis has already been shown in previous works where SVM was used to predict protein-protein interactions [80–82] and the distribution of chemical compounds into drugs and non drugs [83,84]. The interaction data used in this study was obtained from the Adrenergic Receptor (AR) drug and the DrugBank dataset.

The compound feature vector is formed using the mass spectrum as well as the chemical structure. The information of mass spectra is acquired from the NIST 05 mass spectral library [64]. It gives information about the physical-chemical properties of the compounds. The mass-charge ratio (m/z) is used to construct two vectors: fragment vector and gap vector. The various m/z values can be considered as fragments, and the intensity at each of these fragments are used to represent fragment vectors of compounds. The gap vector represents the substructure value at the gap. Here gap is described as the value between two m/z peaks.

The substructure of a chemical compound is considered as a graph, with nodes of the graph representing atoms of the compound. This graphical representation is used to calculate the chemical feature vector using the equation

$$C_l^h(c) = (\varphi_c(p))_{p\varepsilon P_l^h} \tag{6}$$

$\varphi_c(p)$ can be calculated as

$$\varphi_c(p) = \left\{ \begin{array}{cc} \frac{f_c(p)}{\sum_{i\varepsilon P_0^h(c)} f_c(i)} & if \quad p\varepsilon P_l^h(c) \\ 0 & otherwise \end{array} \right\} \tag{7}$$

Here P_l^h is the set of paths of length greater than l and lesser than h and $f_c(p)$ is the frequency of path p in compound c.

The protein feature vector is determined using amino acid sequence. The sequence data is obtained from the UniProtKB knowledgebase [52]. It depicts the basic structure of the protein. The amino acids are connected using peptide bonds that form the primary structure of the protein. To form the vector, the amino acid sequence is first divided into height 1 signatures [85]. Every signature is then mapped to vector space using the formula

$$\alpha_s(s_{01}, s_{11}, s_{12}) = \alpha(s_{01}) + \frac{1}{2}\frac{(\alpha(s_{11}) + \alpha(s_{12}))}{2} \tag{8}$$

Here s_{01} (s_{11}, s_{12}) is the signature and $\alpha(s)$ is a 5D property vector. $\alpha(s)$ is calculated using 237 physical-chemical properties of the amino acids [86]. Bayesian mixture modeling [87] is then performed to cluster these signature vectors into 199 groups. Considering these 199 clusters, the feature vector $F(p)$ is calculated as

$$F(p) = (\beta(c))_{c\epsilon C} \tag{9}$$

$\beta(c)$ is defined as

$$\beta(c) = \left\{ \begin{array}{cc} \frac{f_p(c)}{\sum_{i\varepsilon C(p)} f_p(i)} & if \quad c\varepsilon C(p) \\ 0 & otherwise \end{array} \right\} \tag{10}$$

Here $C(p)$ is the set of clusters and $f_p(c)$ is the frequency of cluster c in protein p. The m dimensional feature vector (v_i) is formed by

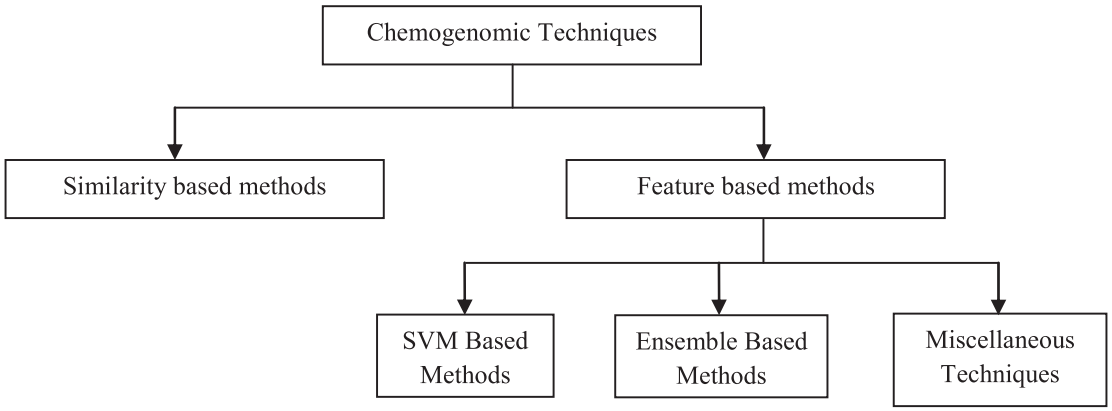

Fig. 3. Flowchart depicting the categorization of chemogenomic techniques.

concatenating the chemical compound feature vector and the protein feature vector to form a single vector. The output ($y_i \varepsilon$ {-1, 1}) which represents the two classes binding and non-binding can be determined using the SVM equation

$$f(v) = sign\ \left(\sum_{i=1}^{n} \alpha_i y_i\ \ K(v_i, v) + b\right) \tag{11}$$

Here v represents the vector of the new object to be classified and $K(.,.)$ is the kernel function. The kernel function used is

$$K_{int}(V_1, V_2) = \prod_{I,J\varepsilon V} K_{IJ(=JI)}(I_1, J_2) \tag{12}$$

$$K_{IJ(=JI)}(x, y) = \begin{Bmatrix} (\gamma_{IJ} x^t y + 1)^3 \\ exp(-\gamma_{IJ}\ ||x - y||^2) \\ \tanh^{(\gamma_{IJ} x^t y + 1)} \\ 1 \end{Bmatrix} \tag{13}$$

Here V is a feature formed by the concatenation of compound and protein features and parameter γ_{IJ} is selected to achieve maximum frequency. To obtain an equalized influence of all vectors, they are normalized and scaled. Although this method achieves good accuracy, it suffers from the limitation that it produces a large number of false positives.

3.1.2. Genome scale enzyme–metabolite and drug–target interaction predictions using the signature molecular descriptor

This technique also uses SVM kernel to predict drug protein interaction [2]. The chemical compounds are depicted using signatures [88]. Canonical sub graphs of these compounds are formed from the molecular graphs which are compared to calculate the similarities between the compounds [89]. Although these similarity scores have been previously used for predicting anti-cancer activities, toxicity etc [84], this study combines them with protein similarities to predict drug target interaction. Protein signatures are also calculated in a similar manner to form the target feature vector. Drug and target features are then classified using SVM kernel.

For every compound as well as protein, molecular signature is calculated. The protein information was taken from the KEGG database [50] and the drug information was obtained from the DrugBank dataset [49]. The molecular signature basically depicts the frequency of occurrence of atomic signature in a molecule [88]. Let the molecular graph of a compound (C) be represented as $G(V,E)$, where V is the set of atoms and E is the set of bonds. The molecular signature of compound C can be calculated as

$$\sigma^h(C) = \sum_{x\varepsilon V} o_G^h(G) \tag{14}$$

Here o_G^h is the unit basis vector of the total number of possible atomic signatures of height h. The chemical compound signature and protein signature are calculated using Eq. (8). The height h of signature used for compounds ranges between 0 and 6 while the height for proteins ranges from 6 to 18. The codes of algorithms to compute the atomic as well as molecular signature have been already explained in previous works [90].

For the SVM classification, the chemical and protein signatures are combined using tenor product. The tensor product of compound C and protein P can be calculated as

$$\sigma^{l,h}(P \otimes\ C) = (p_1c_1, \cdots\cdots p_1c_m, p_2c_1, \cdots\cdots p_2c_m, \cdots\cdots p_nc_1, \cdots\cdots p_nc_m) \tag{15}$$

The SVM uses signature kernel [91] to classify the compound protein pair. Signature kernel is defined as

$$k^h(A, B) = \frac{\sigma^h(A).\ \sigma^h(B)}{|\sigma^h(A)||\sigma^h(B)|} \tag{16}$$

This technique provides one of the pioneering works in the field of drug target interactions and predicts new interactions based on the training dataset consisting of drug protein binding pairs.

3.1.3. Protein–ligand interaction prediction an improved chemogenomics approach

This method also uses SVM for the interaction prediction [3]. However, the difference lies in the fact that it independently computes the compound similarity using a kernel function and the protein similarity using some other kernel function. These kernel similarities are then combined using the tensor product to form the feature vector that can be further processed for classification.

For computing the chemical similarity kernel, Tanimoto kernel [92] is used as it has shown accurate performance in various methods for molecular classification. The kernel can be calculated from the molecule sub graph as

$$K_{chem}(c_1, c_2) = \frac{\varnothing_{chem}(c_1)^T \varnothing_{chem}(c_2)}{\varnothing_{chem}(c_1)^T \varnothing_{chem}(c_1) + \varnothing_{chem}(c_2)^T \varnothing_{chem}(c_2) - \varnothing_{chem}(c_1)^T \varnothing_{chem}(c_2)} \tag{17}$$

Here $\varnothing_{chem}(c)$ is a binary vector. The binary vector is formed of bits, each of which depicts whether the sub graph contains paths of length l or less. The value of l was taken as 8 for good results.

For computing the protein similarity various kernel options were explored. Three kernel functions, namely Dirac kernel [93], Multitask kernel [94] and Hierarchy kernel were proposed. Experiments were conducted using each of these functions to explore which function gives better accuracy. The Dirac kernel basically performs linear classification separately for each target. It can be represented as

$$K_{dirac}(p_1, p_2) = \begin{Bmatrix} 1 & if & p_1 = p_2 \\ 0 & otherwise \end{Bmatrix} \tag{18}$$

The multitask kernel eliminates the orthogonality of the vectors and can be depicted as

$$K_{mutitask}(p_1, p_2) = 1 + K_{dirac}(p_1, p_2) \tag{19}$$

Alternatively, another kernel function is introduced. It includes the hierarchical information of the proteins to calculate the similarity. Three classes of proteins i.e. GPCRs, ion channels and enzymes are used in the study. For enzymes, the Enzyme Commission (EC) number [95] is considered. The GPCRs are divided into four classes based on sequence and functions, which are further subdivided. Voltage dependence and ligand gating are used to classify ion channels. Each of these hierarchies is used to compute the kernel value which can be given as

$$K_{hierarchy}(p_1, p_2) = \{\varnothing_h(p_1), \varnothing_h(p_2)\} \tag{20}$$

Here $\varnothing_h(t)$ represents the number of nodes in the hierarchy for protein p. The tensor product of these drug and target vectors is calculated for further prediction. The tensor product is represented as

$$\varnothing(c, p) = \varnothing_{chem}(c) \otimes\ \varnothing_{tar}(p) \tag{21}$$

This tensor product is then used in the prediction function that depicts whether a protein (p) will bind to a compound (c) or not. The linear function can be represented as

$$f(c, p) = w^T \varnothing(c, p) \tag{22}$$

The sign of function $f(c,p)$ shows whether compound (c) will bind to protein (p) or not. The weight vector w is calculated by training SVM using the interacting and non interacting drug target pairs. The experimental results show that the performance of Multitask and Hierarchy kernel is significantly better than Dirac kernel. This is because Multitask and Hierarchy kernel allow sharing of information across the target class.

3.1.4. Integrating statistical predictions and experimental verifications for enhancing protein-chemical interaction predictions in virtual screening

This technique uses two layer SVM for drug target interaction prediction [96]. The two layer SVM and a novel method of constructing negative dataset are introduced in order to considerably decrease the number of false predictions. Chemical structure information of drugs and protein sequence information is used to form vectors [5]. The chemical feature vector and the protein feature vector are constructed in a similar manner as shown in Section 3.1.1.

After computing the feature vector, 100 first layer SVM models are formed. Each of these models contains random negative samples of compound-protein pairs. Two sets of these 100 SVM models are generated: *allpos* consisting of 1731 positive samples and 1750 negative samples and *subpos* consisting of 534 positive samples and 550 negative samples. The output of first layer SVM is used further in the second layer. The first layer models selected to be used in the second layer are determined by the recursive feature elimination (RFE) method [97]. The second layer uses the DrugBank dataset for positive data and a technique to construct the negative samples [98].

The negative samples are selected by first considering the distances between the positive chemical-protein pairs and all other chemical-protein pairs. These negative samples are then extended based on the following rules:

- min: It contains the first l samples in ascending order of p_i *(iεU-N)*
- max: It contains the first l samples in descending order of p_i *(iεU-N)*
- mle: It contains the first l samples in descending order of p_i *(iεU-N)* and $p_i \leq 0.5$
- mlt: It contains the first l samples in descending order of p_i *(iεU-N)* and $p_i > 0.5$

Here N is the set of tentative negative samples selected and U is the set of chemical-protein pairs except the positive samples and p_i is the probabilistic output of SVM. The second layer SVM uses the RBF kernel [99] for classification which can be depicted as

$$K(a, b) = \exp(-\gamma \, ||a - b||^2) \tag{23}$$

In order to further enhance the performance, feedback strategy has also been proposed. According to this strategy, the sample for the second layer is calculated as

$$S_i = (w \times p_i^a, p_i^1, \cdots\cdots p_i^k) \tag{24}$$

Here p_i^a is the additional feedback model which is trained using feedback data of positive and negative samples, p_i^k is the output of k model of first layer SVM and w is a weight factor.

3.1.5. Analysis of multiple compound–protein interactions reveals novel bioactive molecules

Another SVM based method is Chemical Genomics Based Virtual Screening (CGBVS) that uses molecular patterns to identify drug target interactions and can also identify interactions in novel compounds [7]. The method has been evaluated on GPCRs whose interactions have been obtained from the GLIDA database [56]. For drugs, various chemical properties of the molecular structures, as well as the physiochemical properties, have been taken collectively to form the feature vectors. In order to form the drug feature vector, DRAGONX 1.2 software was used [66]. This software allows the user to load the molecular files and select the various molecular descriptors to be calculated. It automatically calculates these descriptors and their value is stored for further use. The descriptors that have been used in this method include various topological descriptors, functional descriptors, molecular properties etc. The descriptors that depend on the 3D structure of the compounds were not considered. A total of 929 descriptors were calculated. Out of this, the features that showed little or no change across the compounds were eliminated, leaving 797 descriptors to be used further.

The feature set of the proteins was constructed only using the protein sequences. To create features of GPCRs, dipeptide composition-based description was used that constitutes a total of 400 dimensions [100]. For each sub sequence k, this kernel calculates all sub sequences of length k that may differ from the selected sequence by at most m letters. Although these features are simple to understand and construct, they have been shown to be accurate while predicting protein structural classes and sub cellular localizations [101]. The main benefit of using these descriptors is that they can handle sequences of varying length and are alignment free. It was also observed that using various sequence related physiochemical features in addition to these descriptors showed even better accuracy. Thus, 1497 features were calculated using PROFEAT web server [102], that were used in conjunction to other calculated descriptors.

The drug protein feature vectors that are obtained are then concatenated to be classified. The SVM classifier is trained using these features for prediction. Since there is no information available about non interactions, the negative dataset is created by randomly selecting certain pairs that have no interaction information. The number of negative samples generated is equal to the number of positive samples. In addition to identifying whether a drug target pair interacts or not, scores are allocated to each pair based on the confidence level of SVM classification. The Radial Basis Function (RBF) kernel is used for SVM classification. The various parameters of the kernel are optimized using grid search [103]. The RBF kernel for vectors V_1 and V_2 can be represented as

$$K(V_1, V_2) = \exp\left(-\frac{||V_1^2 - V_2^2||^2}{2\sigma^2}\right) \tag{25}$$

Here $||V_1\text{-}V_2||$ represents the Euclidian distance between the two vectors.

3.1.6. Identification of chemogenomic features from drug–target interaction networks using interpretable classifiers

This method proposes a novel technique of extracting drug target interaction information using a limited number of chemogenomic features while maintaining performance and accuracy [9]. The drug target interaction information used for experimental analysis has been retrieved from the DrugBank database [49]. The chemical information of drugs has been taken from the PubChem dataset [51] and the genomic information of targets have been taken from the UniProt [52] and the PFAM datasets [53].

The drugs feature vector is encoded using the PubChem dataset information. 881 dimensional binary feature vector is constructed, where each element represents the presence or absence of PubChem substructure. Similarly, the 876 dimensional protein feature vector is constructed where each element represents the presence or absence of PFAM domain. Let the drug feature vector be represented by $\emptyset(D) = \{d_1, d_2, \ldots d_n\}$ and protein feature vector by $\emptyset(P) = \{p_1, p_2, \ldots p_m\}$. These two vectors are combined using the tensor product as follows:

$$\emptyset(D, P) = (d_1p_1, \cdots\cdots d_1p_m, \cdots\cdots d_np_1, \cdots\cdots d_np_m) \tag{26}$$

These feature vectors are further used to predict drug target interactions using two classifiers: logistic regression [104] and SVM. L1 regularization [105] is used in these classifiers to keep most of the elements of the weight vector equal to non zero values. The SVM optimization can be represented as:

$$\min_w ||w||_1 + C \sum_{i=1}^{n} \log(1 + \exp(-y_i w^T \emptyset(D_i, P_i))) \tag{27}$$

Here $y_i \, \varepsilon \, \{-1, +1\}$. Logistic regression can be represented as:

$$\min_w ||w||_1 + C \sum_{i=1}^{n} \max\{1 - y_i w^T \emptyset(D_i, P_i), 0\} \tag{28}$$

This method is advantageous as compared to kernel based SVM methods as they cannot be used for large scale predictions due to high complexity.

3.1.7. Scalable prediction of compound-protein interactions using minwise hashing

This technique uses minwise hashing [106] procedure in order to generate compact fingerprints for drugs and targets [10]. It significantly reduces the computation time and hence the complexity. The tensor product of the drug compound pairs is used to construct features. Since the tensor product is high dimensional and sparse, minwise hashing can help to optimize the complexity as well as the performance of the technique. The chemical structure of the drugs was obtained from the PubChem database [51] and the protein information was extracted from the UniProt [52] and the PFAM datasets [53]. The drug target interaction information was taken from the STITCH database [58]. The representation of drug and target features as well as the calculation of tensor product is similar to 10.

After calculating the tensor product of the drug target pair, minwise hashing is applied by mapping the feature $\emptyset(D_i T_i)$ to a set S & $\Omega = \{1, 2, \ldots nm\}$. The similarity between the two sets is then calculated using Jaccard coefficient [107]:

$$J(S_a, S_b) = \frac{|S_a \cup S_b|}{|S_a \cap S_b|} \tag{29}$$

Following this, l random permutations are constructed and the resultant string is then projected using the equation:

$$p_{ik} = \min\{\pi_k(S_i)\} \tag{30}$$

Here p_{ik} is the projection and Π_k is the random permutation of set S_i. The minwise hashed values are normally stored using 64 bits. However, to further save space another procedure is proposed. The minwise hashed values are subsequently hashed as

$$s_{ik} = h(p_{ik}) \tag{31}$$

Here h is a random hash function. Thus, the tensor product of the drug target pair is first minwise hashed following which additional hashing is applied to further compact the feature. Finally, the hashed value is expanded to a binary vector $F(D_i, T_i)$. The dimension of this binary vector is much smaller than the original vector. The compact features are classified using linear SVM. Both L1 and L2 regularizations [108] are applied to optimize the weight vector. The L1 and L2 regularization can be represented as

$$\min_w ||w||_1 + \sum_{i=1}^{n} \max\{1 - y_i w^T F(D_i, T_i), 0\} \text{ and}$$

$$\min_w ||w||_2 + \sum_{i=1}^{n} \max\{1 - y_i w^T F(D_i, T_i), 0\} \tag{32}$$

Here $||.||_1$ and $||.||_2$ are L1 and L2 norms respectively. Since the dimension of the compact fingerprint $F(D_i, T_i)$ is much lesser than that of original feature $\emptyset(D_i T_i)$, the complexity of SVM also decreases significantly.

3.2. Ensemble based methods

Ensemble based methods achieve a better performance in comparison to SVM based methods as they use an ensemble of decision trees for training and prediction [109]. The drug and target features that are computed individually are combined and transferred to the ensemble of trees. The resulting outputs are then combined based on statistics like average, geometric mean etc. to compute the final result. The various ensemble based methods have been analyzed in Sections 3.2.1–3.2.5.

3.2.1. A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data

This method uses two important models Random Forest (RF) [110] and SVM for large scale drug target interaction prediction [12]. RF has many advantages. It gives a good performance on large scale multiclass datasets. Moreover, when the number of features is more than the number of samples; it does not cause the problem of overfitting. RF is also highly robust to noisy data. Also, RF uses the bagging procedure for aggregation that reduces the correlation of the trees but preserves their strengths. Thus RF gives good performance and accuracy over high dimensional datasets.

The information about drugs, targets and their interactions was taken from the DrugBank dataset [49]. The chemical descriptors for the drugs were computed using the DRAGON program [66]. A total of 1664 descriptors were calculated using this program. These include various constitutional descriptors, topological descriptors, molecular properties etc. Some of the charge descriptors that could not be computed for all drugs were ignored. Also, some of the descriptors that were almost constant across all the data were excluded. This left a total of 1080 descriptors for further processing. The protein descriptors were formed from the PROFEAT web server [102]. Various descriptors like Dipeptide composition, Moran autocorrelation etc were computed. Various structural and physiochemical properties of proteins were also used to form descriptors. These were added to form the protein features of 1080 dimension.

The experimental dataset was formed by considering all known drug target interactions and forming a positive dataset from them. All other drug target pairs that had no known interactions were used to randomly pick certain pairs and from a negative dataset. After constructing the features as well as the experimental dataset, the RF algorithm is applied. RF is basically a classifier that is formed by combining many decision trees. The output of this classifier is the mode of the outputs of all the trees. Various training sets are formed for individual decision trees and their outputs are calculated. The RF algorithm then computes the combined output. Two parameters are tuned in order to achieve the best performance. These are the number of decision trees to be used and the number of features to make a decision at each node of the tree. Here, Out Of Bag (OOB) error is used to calculate the feature importance [111].

The proposed method successfully identifies drug target interactions for high dimensional data. It can also be used to identify the multi target drugs by recognizing the drugs targeting a group of proteins.

3.2.2. Drug-target interaction prediction via class imbalance-aware ensemble learning

This technique addresses the problem of class imbalance which significantly degrades the drug target prediction performance [14]. The problem is further divided into two issues: between-class imbalance and within-class imbalance. The between-class imbalance refers to the fact that the number of known interacting pairs is considerably smaller than the non interactions. This causes the result to be biased towards the non interacting class [112]. This issue has been dealt with previously by randomly selecting some non interacting pairs and neglecting the other pairs. However, this solution may discard certain important information due to random sampling, which may otherwise help in improving the performance. The other issue, i.e. within-class imbalance, refers to the bias towards the more represented data [113]. The results are expected to be biased towards groups that are represented by more members than those represented by lesser members. This technique provides a solution to both these issues.

The drug target information used in this study has been taken from the DrugBank database [49]. The drug features have been generated using the Rcpi package [73]. This generates various descriptors including topological descriptors, geometrical descriptors etc. Since the sequence of the proteins generates features of variable length, the PROFEAT package [102] was used to generate the protein features. The

features having constant values over the dataset were eliminated. Also, the features that had missing values were replaced by the mean of other values. Subsequently, 193 features for drugs and 1290 features for proteins were considered. The drug target pair feature was formed by concatenating the drug features as well as the protein features. It can be represented as:

$$\varnothing(d, t) = \{d_1, d_2, \cdots\cdots d_{193}, t_1, t_2, \cdots\cdots t_{1290}\} \quad (33)$$

In order to eliminate bias towards any feature, the values of the features are normalized in the range [0,1] using the min-max algorithm. The equations used for normalization are:

$$d_i = \frac{d_i - \min(d_i)}{\max(d_i) - \min(d_i)} \quad (34)$$

$$t_j = \frac{t_j - \min(t_j)}{\max(t_j) - \min(t_j)} \quad (35)$$

After obtaining the normalized features, ensemble learning is used to predict interactions. Ensemble learning involves training each of the base learners using a different training set and then concatenating their results to form the final output. For introducing diversity in the base learners, feature sub spacing is performed. For each of the base learners, two-thirds of the features are used. In order to remove the between-class imbalance, non overlapping sets of negative interactions are considered to be used for training. To eliminate within-class imbalance, oversampling is performed. The known interaction data is divided into K homogeneous clusters. All the clusters thus obtained are oversampled to the size of the largest cluster. Decision trees are trained using the data and the outputs are concatenated to form the result [114].

This method has shown better performance in comparison to the various state of the art methods. It has also been able to identify new drugs for the purpose of drug repositioning.

3.2.3. Drug-target interaction prediction using ensemble learning and dimensionality reduction

This is another method based on ensemble learning to predict drug target interactions. It uses ensemble learning in conjunction with dimensionality reduction to make predictions and develop techniques that can handle the increasing data in the field of drug target interactions. Feature sub spacing is used to introduce diversity just as in Section 3.2.2. For dimensionality reduction, three different methods are used. Finally, the base learners are trained. Two classifiers have been considered for base learners. The dataset used for experimentation has been taken from Section 3.2.2. Also, another dataset is considered which has been taken from Section 3.1.6.

Dimensionality reduction converts the features to a low dimensional space. It reduces the memory requirement as well as the computational complexity. Three dimensionality reduction methods have been considered. They are:

a. SVD

SVD inputs a given matrix $D \varepsilon R^{n \times m}$ and disintegrates into three matrices: $A \varepsilon R^{n \times k}$, $B \varepsilon R^{k \times k}$ and $C \varepsilon R^{m \times k}$ such that $D = ABC^T$. Here B is a diagonal matrix. The values of this matrix are formed by considering the k largest eigenvalues of D. The reduced dimensionality matrix is $D' = AB$. SVD is applied to both drug and target matrices to reduce dimensionality.

b. Partial Least Squares (PLS)

PLS takes two input matrices: predictor variables matrix: $P \varepsilon R^{n \times k}$ and a response matrix: $X \varepsilon R^{n \times m}$ [115]. P and X are first centered and then a model is built using equations:

$$P = UA^T + E \quad (36)$$

$$X = VB^T + F \quad (37)$$

Here U and V are the projections of P and X respectively. E and F are the error terms. To reduce the dimensionality of drug matrix D, P and X are substituted by D and Y, where Y is the interaction matrix. Similarly, for target matrix, P and X are substituted by T and Y^T.

c. Laplacian Eigenmaps

This is a non linear method of dimensionality reduction [116]. For a given matrix $D \varepsilon R^{n \times m}$, the k-nearest neighbor is first formed. If D_i and D_j are neighbors, the graph W can be formed by:

$$W_{ij} = \exp(-||D_i - D_j||^2) \quad (38)$$

Following this, the k-dimensional representation of D (Z) is calculated and an objective function is minimized:

$$Z = argmin_{Z'} \sum_{ij} W_{ij} \, ||Z_i^{'} - Z_j^{'}||^2 \quad (39)$$

After the dimensionality reduction of the drug feature matrix as well as the target feature matrix. Ensemble methods are used. Two ensemble methods used are:

a. Ensemble of decision trees:

Decision trees are used in this model to make predictions. Along with the positive sample, negative data is inputted which is formed by the random sampling of negative data. The drug target feature vector is formed by concatenating the drug vector and the feature vector. Feature sub spacing is applied to introduce diversity in the data. Dimensionality reduction is then applied to the data which is further used for training. The output of the various base learners is averaged to form the final result.

b. Ensemble of Kernel Ridge Regression learners:

This method uses the RLS-avg model for predictions [117]. The sampling of the negative dataset is not performed as this is not required in RLS-avg model. This technique performs dimensionality reduction of the positive and negative dataset and subsequently applies RLS-avg method. RLS-avg computes kernel matrices for both the drug feature matrix as well as the target feature matrix. Following this, the Gaussian Interaction Profile (GIP) is calculated for both drug and target matrices using the drug target interaction matrix. Inner combination model is then used to merge the kernel matrices as well as the GIP kernels. WNN is used in conjunction with RLS-avg to infer interaction profiles of new drugs.

It was concluded that ensemble of Kernel Ridge Regression learners showed better performance as compared to decision trees.

3.2.4. A systematic prediction of drug-target interactions using molecular fingerprints and protein sequences

This technique makes use of extremely randomized trees [118] in order to predict the drug target interactions [20]. The drug feature vector is formed using a novel fingerprint based on the substructure information. For proteins, Pseudo Substitution Matrix Representation (Pseudo-SMR) descriptor is used. Pseudo-SMR retains the biological evolutionary information of the protein sequence. The dataset used in this study is the same as used in a study by Yamanishi et al. [23].

Substructure fingerprints are used to construct drug feature vectors [119]. It fragments the drug and encodes the existence of substructures using a boolean vector. Thus, each bit depicts whether a substructure is present or not. If the substructure is present, it is encoded as 1; otherwise, it is encoded as 0. In this method, the 'PUBCHEM_CACTVS_SUBGRAPHKEYS' property from the PubChem database [51] is used for constructing the 881 dimensional feature vector. To encode the protein

information, Substitution Matrix Representation (SMR) is used. In this method, the BLOSUM62 matrix is used [120]. The matrix can be represented as

$$SMR = \begin{bmatrix} B_{1,1} & B_{1,2} & \cdots & B_{1,20} \\ B_{2,1} & B_{2,2} & \cdots & B_{2,20} \\ \cdots & \cdots & \cdots & \cdots \\ B_{N,1} & B_{N,2} & \cdots & B_{N,20} \end{bmatrix} \quad (40)$$

Here N is the protein sequence length and $B_{i,j}$ depicts whether i^{th} amino acid mutates with j^{th} amino acid. Since this constructs features of variable length, Pseudo SMR is formed to form constant length features [121]. It can be calculated as

$$PSMR(n) = \left\{ \begin{matrix} \frac{1}{N}\sum_{i=1}^{N} P(i,j) n = 1 \cdots 20 \\ \frac{1}{N-lg}\sum_{i=1}^{N-lg} \{P(i,j) - P(i+lg,j)\}^2 j = 1 \cdots 20 \end{matrix} \right\} \quad (41)$$

Here P is the normalized SMR matrix value which can be represented as

$$P(i,j) = \frac{SMR(i,j) - \frac{1}{20}\sum_{a=1}^{20} SMR(i,a)}{\sqrt{\frac{1}{20}\sum_{b=1}^{20}(SMR(i,b) - \frac{1}{20}\sum_{a=1}^{20} SMR(i,a))^2}} \quad (42)$$

The *PSMR(n)* value hence calculated is normalized in the range [0,1]. The drug and target vectors are then fed to Extremely Randomized (ER) trees. These trees are basically an ensemble of regression trees formed from a top down approach. The cut points of these trees are chosen randomly. These ER trees are built by splitting the nodes recursively until the output variables reach a constant value. The splits are evaluated using a score which can be calculated as

$$Sc(s,t) = \frac{2I_s^C(t)}{H_s(t) + H_c(t)} \quad (43)$$

Here $H_c(t)$ is the classification entropy and $H_s(t)$ is the split entropy. $I_s^c(t)$ denotes the mutual information of the split. The various scores are evaluated and the maximum score is chosen as

$$Sc(s^{'},t) = \max_{i=1\cdots K} Sc(s,t) \quad (44)$$

This method gives a good prediction performance. This is mainly due to the fact that the trees are split randomly which helps in reducing bias.

3.2.5. RFDT: A rotation forest-based predictor for predicting drug-target interactions using drug structure and protein sequence information

This technique uses Rotation Forest in order to predict the drug target interactions [25]. The drugs and proteins are encoded in a similar manner as explained in Section 3.2.4. The drugs are encoded using the molecular substructure fingerprints and the proteins are represented by the PSSM matrix. The dataset used in the analysis of the technique is the same as used in a study by Yamanishi et al. [23]. The 881 dimensional drug feature vector is used in conjunction with the target vector. In order to extract the features from the PSSM matrix, Auto Covariance (AC) algorithm is used. It calculates the average correlation between the two protein sequences. The AC value is computed as

$$AC(i,d) = \frac{1}{L-d}\sum_{i=1}^{L-d}(P_{i,j} - \frac{1}{L}\sum_{i=1}^{L} P_{i,j}) \times (P_{i+d,j} - \frac{1}{L}\sum_{i=1}^{L} P_{i,j}) \quad (45)$$

Here d is the distance between the two protein sequences, L is the length of protein sequence and i denotes the amino acid. $P_{i,j}$ is the PSSM matrix value at position *(i,j)*. To construct the protein feature vectors, PSSM has been formed. It is a matrix of N × 20 dimension where N is the length of a particular protein sequence and 20 stands for the number of amino acids. Thus $PSSM_{i,j}$ represents the score of i^{th} amino acid relative to j^{th} position of the protein sequence and can be calculated as

$$PSSM_{i,j} = \sum_{k=1}^{20} p(i,k) \times q(j,k) \quad (46)$$

Here *p(i,k)* represents the frequency of i amino acid appearing at j^{th} position and *q(j,k)* is the value of Dayhoff's matrix of j and k amino acids [122]. The value of d is set by analyzing using fivefold cross validation model.

For classifying the features, an ensemble technique called Rotation Forest (RF) is used. This classifier is based on the entire training data. The entire training feature set is first divided into K subsets of similar size. Principal Component Analysis (PCA) is then applied to each of these subsets. Using the data, L decision trees *(T_1, T_2, … T_L)* are trained. Let n be the total number of features. This suggests that each tree (T_i) contains $F = n/K$ features. The corresponding columns are extracted from the data for the K features to form a new matrix $Y_{i,j}$. Out of this new matrix, 75% of data is chosen to form a new matrix $Y_{i,j}'$. This matrix $Y_{i,j}'$ is used to produce matrix coefficients $M_{i,j}$ These coefficients are used to compute the final rotation matrix R_i as

$$R_i = \begin{bmatrix} \gamma_{i,1}^{1} & \cdots & \gamma_{i,1}^{c1} & 0 & \cdots & 0 \\ 0 & \gamma_{i,2}^{1} & \cdots & \gamma_{i,2}^{c2} & \cdots & 0 \\ & \cdots & \cdots & \cdots & \cdots & \\ 0 & 0 & \cdots & \gamma_{k,1}^{1} & \cdots & \gamma_{k,1}^{k1} \end{bmatrix} \quad (47)$$

This method has been tested on the standard dataset and RF is shown to be more efficient than classification by SVM.

3.3. Miscellaneous techniques

The techniques that do not fall under the above two categories have been compiled under this category. These techniques include various methods for drug target interaction prediction like utilizing Relevance Vector Machines [34], Nearest Neighbour algorithm [27], Sparse Canonical Correspondence Analysis (SCCA) [29] and fuzzy kNN engine [32]. The drug and target features are computed in a similar manner as mentioned in Sections 3.1 and 3.2, following which various machine learning tools are applied to predict interactions. These techniques have been further explained in Sections 3.3.1 to 3.3.5.

3.3.1. Predicting drug-target interaction networks based on functional groups and biological features

This method codes drug feature vectors using functional groups and targets using biological properties [27]. The drug target interaction data is taken from a previous study [23]. The drug target links have been considered for four datasets namely enzymes (E), ion channels (ICs), nuclear receptor (NR) and G protein-coupled receptor (GPCR) targets from different public datasets. These public datasets include KEGG [50], BRITE, BRENDA, SuperTarget [54] and DrugBank [49]. In addition, information about drug and protein properties is obtained from the KEGG database.

Since functional groups represent the properties of a compound, these are used to form the feature vector [123]. A set of 28 functional groups are selected to formulate the feature vector. The vector can be represented as

$$D = [f_1, f_2, \cdots\cdots f_{28}]^T \quad (48)$$

Here f_i depicts whether functional group i is present in drug or not. To formulate the feature vector for proteins, pseudo amino acid composition (Pse-AAC) is proposed [124]. This representation contains the properties of the proteins as well as its sequential information. Six different features were selected for representing proteins. For every feature, various states were defined. E.g.: for feature 'secondary structure' three states were defined: helix, coil and sheet. For every such state, a code is designated. A sequence of all these state codes forms the feature vector. In order to form a fixed length feature vector for all proteins, three properties of the sequence were defined: distribution

(D), composition (C) and transition (T). Here D is the distribution of codes, C is the global composition of each code and T is the frequency of transition of code. These properties are defined for the sequence which results in a 139 D feature vector that can be represented as

$$T = [p_1, p_2, \cdots\cdots\cdots p_{139}]^T \quad (49)$$

Here p_i depicts whether property i is present in the drug or not. Using the drug and target feature, a predictor is constructed with the Nearest Neighbour (NN) algorithm [125]. Since the dataset contains four different classes of proteins, four different predictors are constructed. This algorithm suggests that the drug target pair with the shortest distance with the positive sample is taken as a positive interaction. To measure the nearness of the sample vectors (V_1 and V_2), the following equation was used

$$D(V_1, V_2) = 1 - \frac{V_1 . V_2}{||V_1|| ||V_2||} \quad (50)$$

To improve the performance and efficiency of the technique, only those features that have an impact on the prediction performance are included. The redundant features are excluded using the Maximum Relevance Minimum Redundancy (mRMR) algorithm [126]. For this, the mutual information MI of two features f_1 and f_2 are calculated

$$MI(f_1, f_2) = \iint p(f_1, f_2) \ log \frac{p(f_1, f_2)}{p(f_1)p(f_2)} dx \ dy \quad (51)$$

Following this, the mRMR function is used to rank the relevance of the features. The features form an ordered list according to their importance. The mRMR function used is

$$\max_{f_j \in \delta_t} [MI(f_j, c) - \frac{1}{m} \sum_{f_i \in \delta_s} MI(f_j, f_i)] \quad (52)$$

Here δ_s is the feature set that has already been chosen and δ_t is the set of features to be selected. The ordered list of features that has been formed is further processed using Incremental Feature Selection (IFS) to determine which features need to be included for prediction. The features are added one by one according to their importance to form N different feature sets. A feature set F_i can be shown as

$$F_i = [f_1, f_2, \cdots\cdots\cdots f_i] \quad (53)$$

Using each of these N feature sets, the prediction is performed and tested using Jackknife cross validation to determine the feature set for optimum performance. To refine the selection of optimum feature set selection Forward Feature Selection (FFS) is performed. The predictor is run for each feature set and the feature set for which the accuracy reaches its peak is selected iteratively using FFS.

3.3.2. Extracting sets of chemical substructures and protein domains governing drug-target interactions

This technique identifies drug target interactions using SCCA, an extension of Canonical Correspondence Analysis (CCA) [127], for drugs as well as targets simultaneously [29]. The proteins binding to a common set of chemicals are clustered together for further prediction process. The chemical substructure information from the PubChem database [51] is used in conjunction with genomic information and annotation of proteins from the UniProt [52] and the PFAM database [53] to construct drug and protein features respectively.

In CCA, the canonical correlation coefficient is represented as

$$corr(u, v) = \frac{\sum_{i,j} I(a_i, b_j) \alpha^T a_i . \beta^T b_j}{\sqrt{\sum_i d_{a_i} (\alpha^T a_i)^2} \sqrt{\sum_j d_{b_j} (\beta^T b_j)^2}} \quad (54)$$

Here we consider n_a drugs and n_b proteins with p and q number of features respectively. u and v are the canonical components of a and b respectively and α and β are the weight vectors for a and b respectively. $I(.,.)$ represents the indicator function and d represents the degree. Further, the adjacency matrix X is constructed where $X_{ij} = 1$ if drug a_i interacts with protein b_j. A and B are the drug feature and protein feature matrices respectively. These are used to define the maximization problem

$$\max\{\alpha^T A^T XB\beta\} \ \ subject \ \ to \ \ \alpha^T A^T D_a A\alpha \leq 1,$$

$$\beta^T B^T D_b B\beta \leq 1 \quad (55)$$

Here D_a and D_b are the diagonal matrices where the diagonals are the degrees of drugs and targets respectively. Consequently, SCCA maximization problem can be represented as

$$\max\{\alpha^T A^T XB\beta\} \ \ subject \ \ to \ \ ||\alpha||_2^2 \leq 1, ||\beta||_2^2 \leq 1, ||\alpha||_1 \leq c_1 p, ||\beta||_1 \leq c_2 q \quad (56)$$

Here $||.||$ is the L_1 norm and parameters c_1 and c_2 control the sparsity of matrices. In order to solve the above equation, penalized matrix decomposition (PMD) [128] can be used. The PMD algorithm can be used to find one canonical component In order to find multiple canonical components, this algorithm can be iterated repeatedly. The SCCA algorithm is used to compute the prediction score for a drug (a) - target (b) pair. The score can be calculated as

$$s(a, b) = \sum_{k=1}^{m} a^T \alpha_k \rho_k \beta_k^T y \quad (57)$$

Here m is the total number of canonical components and ρ_k represents the k^{th} singular value. If the prediction score is greater than a threshold value, drug a is predicted to interact with protein b.

This method can be used in the process of drug discovery as well as to identify the off-target proteins. The major limitation of this technique is the difficulty in choosing the sparsity controlling parameters as well as the number of canonical components to be considered.

3.3.3. Combining drug and gene similarity measures for drug-target elucidation

New technique Similarity-based Inference of drug-TARgets (SITAR) was proposed that introduced many novel similarity measures and integrated these measures to predict drug target interactions [31]. The method uses five drug-drug similarity measures and three gene-gene similarity measures to overcome the limitations of using a single similarity measure.

The drug-drug similarity measures used are:

a. Chemical based: The similarity was computed by first acquiring the simplified molecular input line entry specification (SMILES) of drugs from DrugBank [49]. Subsequently, hashed fingerprints were calculated using the Chemical Development Kit (CDK) [67] and the similarity score was computed using 2D Tanimoto score [129].
b. Ligand based: For ligand based similarity, Similarity Ensemble Approach (SEA) [130] search tool compares SMILES to ligand sets in order to compute the E values. The drug ligand pairs with E values $\geq 10^{-5}$ were considered further and Jaccard score [131] was calculated for them.
c. Expression based: The Connectivity Map Project [132] was used to represent the gene expression profiles. These were ranked using the Spearman Rank coefficient, the Jaccard coefficient and a method proposed by Iorio et al. [133].
d. Side Effect based: The drug side effects with chemical properties were used to calculate similarity [134]. Jaccard coefficient was then employed to calculate similarity based on top ten predicted side effects.
e. Annotation based: Anatomical Therapeutic Chemical (ATC) classification system [135] was used to assign ATC codes to drugs. The Semantic similarity algorithm [136] was employed to calculate the similarity.

The gene-gene similarity measures used are:

a. Sequence based: Smith Waterman alignment scores [137] were used to calculate the sequence similarities of the proteins.
b. Closeness in network: Protein-protein interaction network was constructed [138–142] and the distances between the proteins were computed using the All pairs shortest path algorithm. These distance values were then converted to similarity scores.
c. Gene Ontology (GO) semantic similarity: Semantic similarity was computed using the R package [143] by utilizing the GO annotations [144] from the UniProt database.

The drug target interaction information was extracted from KEGG DRUG [50], DrugBank [49], and DCDB [145] and Matador [55]. The features were formed by considering a combination of one of the drug-drug similarity measures and one of the gene-gene similarity measures. These two similarities were integrated to form a single score. The similarities are combined using a logistic regression classifier with a wrapper function. To calculate the score, the maximum value of the geometric mean was considered:

$$Score(d, g) = \max_{d,g \neq d,g} S(d, d')^r . S(g, g')^{(1-r)} \quad (58)$$

Here *S(d,d')* is the drug-drug similarity, *S(g,g')* is the gene-gene similarity and r is the scoring parameter.

3.3.4. iGPCR-drug: A web server for predicting interaction between GPCRs and drugs in cellular networking

This technique primarily focuses on the prediction of interactions between chemical compounds and a specific class of proteins i.e. GPCRs [32]. Since GPCRs frequently serve as targets for most therapeutic drugs [146], they have been the focus of study in this method. This technique also introduces an open source web server called iGPCR for public usage to increase the practicability of interaction prediction.

In order to generate the drug feature vectors, the MOL files of drugs from the KEGG database [50] are considered. This MOL file is then fed to a software toolbox called OpenBabel [68] to generate molecular fingerprints of the drugs. It identifies small molecules from the drugs and converts them to bit-string of 256 bits using a hash function. Let V_1, V_2, … V_{256} be the bits of the vector. The Fourier transform of this representation is then computed to represent the inwardness by

$$F_k = \sum_{i=1}^{256} V_i \exp\left[-j\left(\frac{2\pi l}{256}\right)k\right] \quad (59)$$

For the computed Fourier transform values, the amplitude is calculated as

$$A_k = [real^2(F_k) + imag^2(F_k)]^{1/2} \quad (60)$$

Using these values the drug feature vector is formed as

$$D_i = [A_1, A_2 \cdots\cdots A_{256}]^T \quad (61)$$

For representing the protein vectors, Pse-AAC composition is used. The protein vector can be depicted as

$$T_i = [\psi_1, \varphi_2, \cdots\cdots \psi_n]^T \quad (62)$$

Here, the components ψ_i depend on the desired information that has been extracted from the protein sequences. The information used in this method is the mean polarity value of the various amino acids. The drug and the target feature vectors are then combined using the orthogonal sum. The combined feature vector can be shown as

$$X = T \oplus wD = [\psi_1, \psi_2, \cdots\cdots \psi_{22}, wA_1, wA_2, \cdots\cdots wA_{256}] \quad (63)$$

Here w is the weight vector that has been set as *1/700* in this method. The feature vector is then transferred to a fuzzy kNN engine [125] due to its efficiency while dealing with complex biological systems [147–149]. Let *[F_1, F_2, …F_n]* be the query pairs and *[F_1*, F_2*, … F_K*]* be the neighbors of the query pairs, the fuzzy membership function can be defined as

$$\mu^+(X) = \frac{\sum_{j=1}^{k} \mu^+(F_j^*) d(F, F_j^*)^{-2/(\varphi-1)}}{\sum_{j=1}^{k} d(F, F_j^*)^{-2/(\varphi-1)}} \quad (64)$$

$$\mu^-(X) = \frac{\sum_{j=1}^{k} \mu^-(F_j^*) d(F, F_j^*)^{-2/(\varphi-1)}}{\sum_{j=1}^{k} d(F, F_j^*)^{-2/(\varphi-1)}} \quad (65)$$

Here *d(.,.)* is the Euclidean distance and ψ is the fuzzy coefficient. The predictor thus formed has been named as iGPCR-Drug and is freely available online.

3.3.5. Prediction of drug–target interaction networks from the integration of protein sequences and drug chemical structures

This method, named as PDTPS (Predicting Drug Targets with Protein Sequence) is used to predict the drug target interactions using the Bigram Probabilities (BIGP) of PSSM for proteins [34]. This method uses Relevance Vector Machines (RVM) [150] for the classification purpose. The data used for the experimental analysis of this technique has been extracted from DrugBank [49], BRENDA [151], KEGG BRITE [50] and SuperTarget [54]. The data has been previously used in a study by Yamanishi et al [23].

In order to construct the PSSM matrix, Position Specific Iterated BLAST (PSI-BLAST) was used [152]. An e-value of 0.001 was set and three iterations were considered in PSI-BLAST. The PSSM matrix is then used to calculate the Bigram Probabilities. BIGP feature extraction method has been used to calculate these probabilities. $P_{i,j}$ is considered as the relative probability of j^{th} amino acid to occur at i^{th} location of the protein sequence. It can be represented as

$$P_{i,j} = \sum_{j=1}^{20} i{:}\ 1 = 1\cdots\cdots N,\ j = 1\cdots\cdots 20 \quad (66)$$

The bigram probability $BIGP_{mn}$ for m and n amino acid can then be shown as

$$BIGP_{mn} = \sum_{i=1}^{L-1} P_{i,m} P_{i+1,n} 1 \leq m \leq 20,\ 1 \leq n \leq 20 \quad (67)$$

The matrix thus formed is called as bigram occurrence matrix and can be represented as

$$BIGP = \{BIGP_{1,1}, \cdots\cdots BIGP_{1,20}, BIGP_{2,1}, \cdots\cdots BIGP_{20,20}] \quad (68)$$

Thus a 400 dimensional vector was formed from each of the protein sequences. In order to reduce the dimension as well as the noise, PCA was employed [153]. The dimension was reduced from 400 to 350. The bigram probabilities representation of the proteins as well as the chemical structure representation of the drugs are used to construct the drug target pair feature vector. The training set is then used to train the classifier i.e. RVM. RVMs use Bayesian inference in order to classify the data. The model of RVM can be represented as

$$k(x, x') = \sum_{j=1}^{n} \frac{1}{\alpha_j} \varphi(x, x_j) \varphi(x', x_j) \quad (69)$$

Here X represents the training data, ψ is the kernel function and α_j depicts the variances of the weight vector.

This technique has shown a good performance over the standard dataset in comparison to the SVM classifier. The improvement of this method over the past techniques is that it uses PSSM which retains the evolutionary information and helps in improving efficiency. Also, BIGP reduces the sparsity which increases the performance. The comparison shows the efficiency as well as the robustness of the method.

4. Comparative analysis of various feature based techniques

The methods and techniques discussed in Section 3 predict the drug target interactions by using various classifiers and features. The details of these techniques have been compiled in Table 2. It summarizes the

Table 2
A summary of feature based techniques.

Reference no.	Drug features	Drug dataset	Target features	Target dataset	Technique/classifier	Drug target interaction dataset
[1]	Mass spectrum and chemical structure	NIST 05 mass spectral library	Amino acid sequence	UniProtKB knowledgebase	SVM	Adrenergic Receptor (AR) drug and DrugBank
[2]	Molecular signature	DrugBank	Molecular signature	KEGG	SVM	DrugBank
[3]	Chemical structure	KEGG BRITE	Hierarchical information	KEGG	SVM	KEGG BRITE
[5]	Chemical structure	DrugBank	Amino acid sequence	DrugBank	Two layer SVM	DrugBank
[7]	Chemical and physiochemical properties	GLIDA and GVK Biosciences kinase inhibitor database	Sequence based properties	GLIDA and GVK Biosciences kinase inhibitor database	SVM	GLIDA and GVK Biosciences kinase inhibitor database
[9]	Chemical structure	PubChem database	Sequence based properties	UniProt and PFAM database	SVM and logistic regression	DrugBank
[10]	Chemical structure	PubChem database	Sequence based properties	UniProt and PFAM database	SVM	STITCH
[12]	Chemical and physical properties	DrugBank	Protein properties	DrugBank	Random Forest	DrugBank
[14]	Chemical and physiochemical properties	DrugBank	Sequence based properties	DrugBank	Ensemble learning	DrugBank
[19]	Chemical structure, chemical and physiochemical properties	DrugBank and PubChem database	Sequence based properties	DrugBank, UniProt and PFAM database	Ensemble learning	DrugBank
[20]	Chemical substructure fingerprints	PubChem	SMR matrix	KEGG GENES	Extremely Randomized (ER) trees	Dataset by Yamanishi et al [23]
[25]	Substructure information	DrugBank	Bigram probabilities from PSSM matrix	KEGG GENES	Random Forest (RF)	Dataset by Yamanishi et al [23]
[27]	Functional groups	KEGG	Biological features	KEGG	Nearest Neighbour predictor	Dataset by Yamanishi et al [23]
[29]	Chemical substructure	PubChem database	Genomic information and annotation	UniProt and PFAM database	SCCA	DrugBank
[31]	Chemical, ligand, expression, side effect and annotation based	DrugBank and SIDER database	Sequence, network and gene antology based	KEGG and UniProt database	Logistic regression classifier	KEGG DRUG, DrugBank, and DCDB and Matador
[32]	Molecular fingerprints	KEGG	Biological features	KEGG	Fuzzy kNN engine	KEGG
[34]	Substructure information	DrugBank	Bigram probabilities from PSSM matrix	KEGG GENES	Relevance Vector Machines (RVM)	Dataset by Yamanishi et al [23]

drug and target features used in each of the techniques along with the datasets used to construct these features. It also mentions the dataset for retrieving the drug target interaction information as well as the classifier used in each of the techniques.

The feature based methods have several advantages. They provide a straightforward approach to predict the drug target interactions. The features represent the characteristics of the drugs as well as the targets. This preserves the data characteristics that can be used for interpretability. They provide an efficient mechanism of drug target interaction prediction with accurate results. Also, various classifiers, as well as ensemble of classifiers, have been proposed in the field of machine learning that can be employed for the process of classification directly. The ongoing research in the field of classification and machine learning will greatly benefit the exploration of the feature based methods.

As mentioned in Section 2.3, many metrics have been proposed to compare and contrast the various techniques o drug target interaction prediction. However, a major drawback is that most of the techniques use a different dataset for evaluation of results. As such, it is not possible to quantitatively compare the methods based on accuracy, precision etc as different datasets may produce different results. The lack of a standard dataset does not allow the application of statistical measures to compare the techniques. However, each of the methods has certain advantages as well as limitations that can provide an overview in terms of the comparison of the techniques. The SVM based methods are simple to comprehend as well as execute. SVM also produces satisfactory results on different kinds structured or semi-structured of data. By selecting an appropriate kernel function, SVM produces efficient results. SVM classifier proves to be a good choice for high dimensional data and also prevents overfitting [154]. Section 3.1 outlines the various SVM based methods as well as the different kernel functions that have been used for drug target interaction prediction. Apart from the SVM, certain other techniques have also been proposed. Nearest neighbor method has been used in conjunction with MRMR and IFS to select optimal features [27]. The application of Sparse Canonical Correspondence Analysis has also been proposed to form clusters of similar chemical structures and protein domains to facilitate the interaction prediction [29]. A web server using fuzzy kNN has also been developed specifically for GPCR family of proteins to predict their interactions [32]. Relevance Vector Machines have also been used with PCA to acquire reduced dimensional features and provide a low complexity method [34]. Additionally, different drug and target similarity measures have been explored and used in combination to identify an optimal combination of these measures to improve performance [31]. Although some of these methods perform better than the traditional SVM techniques, they do not provide any significant improvement.

Ensemble based methods, on the other hand, have higher accuracy and performance in comparison to SVM and miscellaneous techniques. Ensemble based methods aggregate the results of various classifiers in order to improve accuracy. Thus, ensemble methods work better than using just a single model like SVM. Methods like Random Forest [12] and Extremely Randomized Trees [20] introduce randomness in the training set that produces better results than classifiers like SVM, ANN etc.

A comparison of various drug and target features also shows that the chemical structure and physiochemical features of the drugs give better results. The structure of a drug directly influences its activity with the target protein. The physiochemical properties also describe the chemical as well as the physical information of a drug. Thus, they help to identify the drug reactions and their binding capability to a target. As for proteins, the use of sequence information has shown better performance. The sequence information depicts the amino acid linear sequencing within a protein. The sequence information of a protein greatly determines their ability to bind to a specific chemical. Hence, it provides a distinguishing feature that helps in drug target interaction perdition.

The feature based methods however also suffer from certain limitations. The feature based methods have greater simulation runtime in comparison to the similarity based methods. The process of calculating and extracting features is time consuming which increases the complexity. Moreover, the process of selecting and extracting optimal features is a complex task. The features calculated for the drugs and the targets need to be evaluated to find the optimal set of features that provide greater accuracy. But, various feature selection and feature extraction methods are complex adding to the complexity of the method.

5. Discussion

The feature based methods have been categorized into SVM based methods, ensemble based methods and miscellaneous techniques. The problem of drug target interaction prediction was first addressed using SVM classifiers to predict the interaction/non interaction of the drug target pair. As mentioned in Section 4, SVM is easier to implement and produces satisfactory results on high dimensional structured and semi structured data. The application of two layer SVM has also significantly improved the prediction performance [5]. However, it has certain limitations. Selecting an appropriate kernel function based on the data is a complex task. Also, various parameters need to be tried and tested for the SVM to produce optimal results. Training and testing by SVM are also time consuming. With the development of more advanced machine learning techniques like ensemble learning, relevance vector machines etc, more efficient algorithms have been proposed to predict drug target interaction as well. The development of the field of machine learning will coincidently enhance the research in the field of drug target interaction prediction.

The various miscellaneous techniques have also aimed to improve the efficiency of drug target interaction prediction and have established various important insights regarding the drug and target features and their effect on interaction prediction. For instance, Pse-AAC was used to construct target features [27]. It depicts the protein sequence with the help of a model but also preserves its sequence information. Similarly, SCCA was proposed to form drug and target features [29]. It is a multivariate method to explain the relationship between the data. Although SCCA performs quite well and is immune to noise and skewed data, it also suffers from certain limitations. It is difficult to optimize the various parameters and components for an efficient performance using SCCA. Also SCCA might not produce satisfactory results in the cases where the data is highly correlated and sparsity is not a vital characteristic of drug and target descriptors. Various similarity measures for drugs and targets have also been explored. Drug-drug similarity measures like chemical based, ligand based etc, and target-target similarity measures like sequence based, gene ontology based etc were studied in order to find the best combination for predicting drug target interactions [31]. It has been shown that the combination of ligand based drug similarity and sequence based target similarity has produced the best AUPR scores. Different methods of classification like the fuzzy kNN engine [32], RVM [34], nearest neighbor [27] etc. have also been employed. Though these techniques have not offered a substantial improvement in terms of accuracy, they can be used as a base model to develop more efficient techniques.

Ensemble based methods also efficiently handle large datasets and large number of features representing the drugs and the targets. Rotation forest, for instance, utilizes the entire dataset efficiently and produces different results by rotating the axis of the tree that introduces diversity in the results [25]. A comparative analysis shows that rotation forest produces more accurate and precise results in comparison to SVM. Random forest also aims at improving the efficiency of drug target interaction prediction [12]. It efficiently handles large datasets and is immune to overfitting. However, it is a kind of a black box technique to introduce randomness and produces results that are not interpretable. Another ensemble method, known as extremely randomized trees, has also been proposed [20]. They are computationally inexpensive in

comparison to random forest and rotation forest and produce better results than certain similarity based methods. However, their accuracy is still less in comparison to the state of the art feature based methods. Along with improving efficiency, ensemble learning also aims to address the problem of class imbalance. The within-class imbalance and the between-class imbalance are significantly reduced using class imbalance aware ensemble learning methods [14]. The between-class imbalance refers to the bias towards better represented classes and the within-class imbalance refers to the bias towards non interacting drug target pairs. Dimensionality reduction is also introduced along with ensemble learning to reduce the running time and complexity [19]. The evaluative results show that this method has better performance in comparison to other methods like SVM, RF and various similarity based methods.

There are however many future research directions that can improve the performance as well as the accuracy of the feature based methods. The ensemble based methods can be explored further by using an aggregation of other classifiers and evaluating their accuracy. Different kernel functions can also be analyzed to improve performance. In order to reduce the complexity, feature selection and feature extraction methods can also be examined. Various dimensionality reduction techniques can significantly reduce the complexity of the methods. Many novel techniques are being developed in order to identify the disease genes [155–158]. The relationship between disease genes and the drug targets has not been explored much. The information of disease genes can be incorporated in the drug target interaction process for making predictions.

6. Proposed framework

Ensemble based methods for the prediction of drug target interaction have proven to be more accurate than the traditional methods using a single classifier. These methods have been inspired by the natural functioning of the brain. The modular functioning divides each of the tasks into sub tasks and processes them individually [159]. Thus, it helps to divide a problem into smaller elements to process them without any mutual interference. A novel technique has been proposed in this section using ensemble based classifiers that is expected to deliver better performance than the existing state of the art techniques. The basic steps of the technique have been depicted in Fig. 4.

According to the literature survey and comparison, the chemical and physiochemical properties of the drugs have shown better performance. For targets, the sequence based properties have given better accuracy. Thus, we propose the use of chemical and physiochemical descriptors of the drugs and the sequence based descriptors of the proteins to construct the feature vectors. The drug and target features are then combined to form a single feature vector for a drug target pair. The final feature vectors can be depicted as

$$F = [f_{11}, f_{12}, \cdots\cdots f_{1n}, f_{21}, \cdots\cdots f_{2n}, \cdots\cdots f_{m1}, \cdots\cdots f_{mn}] \quad (70)$$

Here n is the total number of drugs and m is the total number of targets.

These features are then used to train an ensemble of classifiers. Since the dimensionality of the features is very large, they increase the complexity of the system. Dimensionality reduction has been proposed as a feasible solution to decrease the complexity of the methods [19,25]. Thus dimensionality reduction will be applied to the feature vectors. PCA has been proposed as a dimensionality reduction method [153]. PCA reduces the vector dimensionality to down sample the high dimensional data without losing important information.

An ensemble of classifiers has been proposed to increase the accuracy of prediction and reduce the false negatives produced during the testing phase. The existing ensemble based methods train the similar classifier over subsets of the data. However, this does not significantly decrease the false negatives. Let us consider an ensemble of similar classifiers: $[C_1, C_2, \ldots C_p]$. If the set of classifiers are identical, the wrong results will be predicted incorrectly in all of the base classifiers. Hence, the aggregated result will still contain incorrect predictions. However, if the classifiers are diverse, the errors are not related to each other. In case a classifier C_i makes an incorrect prediction, most of the other classifiers may produce a correct result. Thus, the aggregation of these individual results leads to better performance.

Since many SVM based methods have been proposed for drug target interaction prediction, an ensemble of SVM classifiers is proposed to achieve better results. A single SVM classifier does not always produce globally optimum results over the entire dataset. Thus, an ensemble of SVM classifiers may improve the performance of the technique. In order to construct the training set for the ensemble of classifiers, the boosting technique is followed. According to the boosting procedure, an initial training set is considered. Let $[x_1, x_2, \ldots x_l]$ be the initial training set. A weight is assigned to each of these features i.e. $w(x_i) = 1/l$. For the k^{th} classifier, a dataset $[x_1, x_2, \ldots x_{l'}]$ is constructed based on the weight

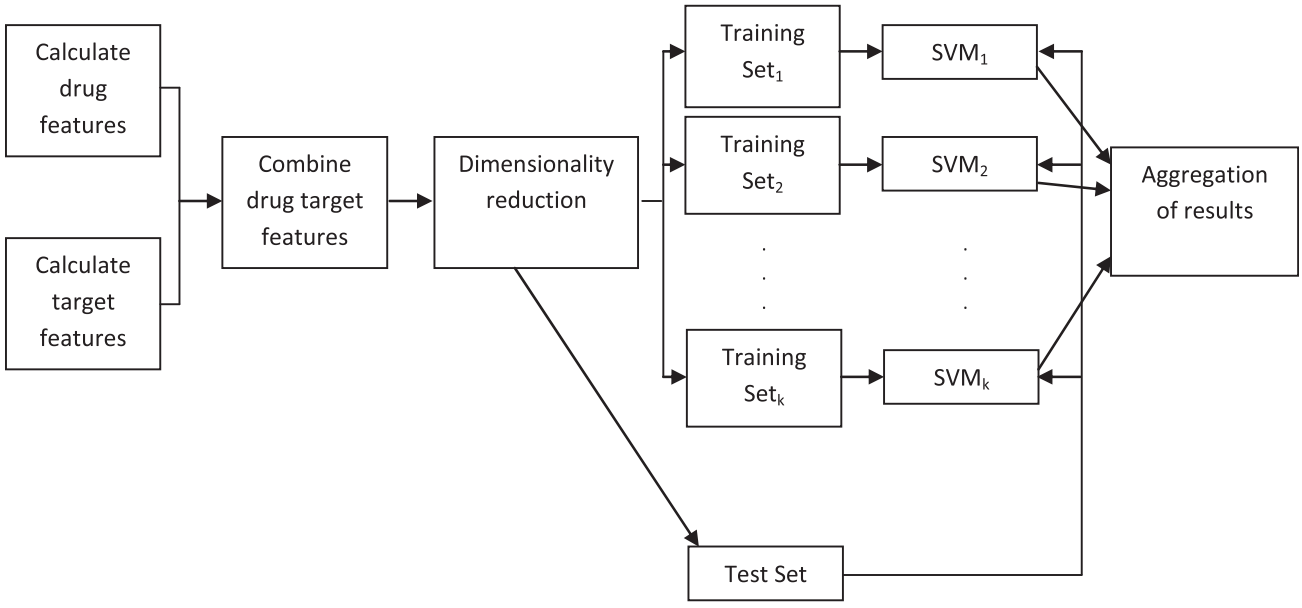

Fig. 4. Steps of proposed framework.

values $w_{k-1}(x_i)$. This set is used to train the k^{th} SVM classifier. The performance of the classifier is then tested using the entire dataset and the weight values are updated. The weight of the predictions which are incorrect is increased. Also, the weight of correctly classified samples is decreased. This implies that the samples which are difficult for classification are used more frequently to train the classifier. This sampling and weight updating are repeated k times to build the SVM classifier.

In order to combine the results of the SVMs, two techniques are proposed: majority voting and hierarchical combination [109]. Majority voting is a simple linear aggregation method. According to this method, all the outputs of the various classifiers are considered. Each output is considered as a vote. The output with more than half of the votes is considered as the final result. The hierarchical method, on the other hand, is a non linear process. The outputs of the lower layer SVM are fed to the upper layer SVMs. This forms a hierarchical structure. Let the output functions of the k SVMs be $f_1, f_2 \ldots f_k$ respectively. The final decision function F corresponds to the output function of the upper layer SVM. The result for a test vector X can be depicted as

$$f_{final} = F(f_1(x), f_2(x), \cdots\cdots f_k(x)) \quad (71)$$

The result of the ensemble of classifiers is considered as the final prediction. The diverse classifiers will help to achieve better accuracy.

7. Conclusion

In this paper, we aim to compile the important contributions in the field of feature based techniques for drug target interaction prediction. It presents an up to date review of the various feature based methods. It also explains the related datasets, tools for calculating drug or target features and the metrics for evaluation. An analysis section intends to compare the performance of various methods outlining their advantages and disadvantages. A novel framework has also been presented which is expected to increase the efficiency of drug target interaction prediction.

Drug target interaction prediction has been one of the important fields of research in drug discovery. It has been continuing since almost a decade and aims to continuously improve the performance of methods with various technologies, features and frameworks. Novel datasets further enhance the performance of prediction. More sophisticated algorithms and the use of big data techniques may further contribute to the research in this area.

Declaration of interest

The authors declare that there are no conflict of interest.

References

[1] N. Nagamine, Y. Sakakibara, Statistical prediction of protein–chemical interactions based on chemical structure and mass spectrometry data, Bioinformatics 23 (15) (2007) 2004–2012.
[2] J.-L. Faulon, et al., Genome scale enzyme–metabolite and drug–target interaction predictions using the signature molecular descriptor, Bioinformatics 24 (2) (2007) 225–233.
[3] L. Jacob, J.-P. Vert, Protein-ligand interaction prediction: an improved chemogenomics approach, Bioinformatics 24 (19) (2008) 2149–2156.
[4] T. Takenaka, Classical vs reverse pharmacology in drug discovery, BJU Int. 88 (2001) 7–10.
[5] N. Nagamine, et al., Integrating statistical predictions and experimental verifications for enhancing protein-chemical interaction predictions in virtual screening, PLoS Comput. Biol. 5 (6) (2009) e1000397.
[6] A. Ezzat, et al., Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey, Briefings Bioinf. (2018) bby002.
[7] H. Yabuuchi, et al., Analysis of multiple compound–protein interactions reveals novel bioactive molecules, Mol. Syst. Biol. 7 (1) (2011) 472.
[8] K. Thongprasom, et al., Interventions for treating oral lichen planus, Cochrane Database Syst. Rev. 7 (7) (2011).
[9] Y. Tabei, et al., Identification of chemogenomic features from drug–target interaction networks using interpretable classifiers, Bioinformatics 28 (18) (2012) i487–i494.
[10] Y. Tabei, Y. Yamanishi, Scalable prediction of compound-protein interactions using minwise hashing, BMC Syst. Biol. 7 (6) (2013) S3.
[11] D.M. Kay, et al., Parkinson's disease and LRRK2: frequency of a common mutation in US movement disorder clinics, Movement Disorders: Off. J. Movement Disorder Soc. 21 (4) (2006) 519–523.
[12] H. Yu, et al., A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data, PloS One 7 (5) (2012) e37608.
[13] X. Chen, et al., Drug–target interaction prediction: databases, web servers and computational models, Briefings Bioinf. 17 (4) (2016) 696–712.
[14] A. Ezzat, et al., Drug-target interaction prediction via class imbalance-aware ensemble learning, BMC Bioinf. 17 (19) (2016) 509.
[15] A.L. Hopkins, Drug discovery: predicting promiscuity, Nature 462 (7270) (2009) 167.
[16] J.T. Dudley, T. Deshpande, A.J. Butte, Exploiting drug–disease relationships for computational drug repositioning, Briefings Bioinf. 12 (4) (2011) 303–311.
[17] S.J. Swamidass, Mining small-molecule screens to repurpose drugs, Briefings Bioinf. 12 (4) (2011) 327–335.
[18] F. Moriaud, et al., Identify drug repurposing candidates by mining the Protein Data Bank, Briefings Bioinf. 12 (4) (2011) 336–340.
[19] A. Ezzat, et al., Drug-target interaction prediction using ensemble learning and dimensionality reduction, Methods 129 (2017) 81–88.
[20] Y.-A. Huang, Z.-H. You, X. Chen, A systematic prediction of drug-target interactions using molecular fingerprints and protein sequences, Curr. Protein Pept. Sci. 19 (5) (2018) 468–478.
[21] E. Lounkine, et al., Large-scale prediction and testing of drug activity on side-effect targets, Nature 486 (7403) (2012) 361.
[22] E. Pauwels, V. Stoven, Y. Yamanishi, Predicting drug side-effect profiles: a chemical fragment-based approach, BMC Bioinf. 12 (1) (2011) 169.
[23] Y. Yamanishi, et al., Prediction of drug–target interaction networks from the integration of chemical and genomic spaces, Bioinformatics 24 (13) (2008) i232–i240.
[24] L. Yao, J.A. Evans, A. Rzhetsky, Novel opportunities for computational biology and sociology in drug discovery: corrected paper, Trends Biotechnol. 28 (4) (2010) 161–170.
[25] L. Wang, et al., Rfdt: A rotation forest-based predictor for predicting drug-target interactions using drug structure and protein sequence information, Curr. Protein Pept. Sci. (2016).
[26] S.M. Paul, et al., How to improve R&D productivity: the pharmaceutical industry's grand challenge, Nat. Rev. Drug Discovery 9 (3) (2010) 203.
[27] Z. He, et al., Predicting drug-target interaction networks based on functional groups and biological features, PloS One 5 (3) (2010) e9603.
[28] H. Chen, Z. Zhang, A semi-supervised method for drug-target interaction prediction with consistency in networks, PloS One 8 (5) (2013) e62975.
[29] Y. Yamanishi, et al., Extracting sets of chemical substructures and protein domains governing drug-target interactions, J. Chem. Inf. Model. 51 (5) (2011) 1183–1194.
[30] J.-J. Lu, et al., Multi-target drugs: the trend of drug research and development, PloS One 7 (6) (2012) e40262.
[31] L. Perlman, et al., Combining drug and gene similarity measures for drug-target elucidation, J. Comput. Biol. 18 (2) (2011) 133–145.
[32] X. Xiao, et al., iGPCR-Drug: a web server for predicting interaction between GPCRs and drugs in cellular networking, PloS One 8 (8) (2013) e72234.
[33] A. Frolov, et al., Response markers and the molecular mechanisms of action of gleevec in gastrointestinal stromal tumors1, Mol. Cancer Ther. 2 (8) (2003) 699–709.
[34] F.-R. Meng, et al., Prediction of drug-target interaction networks from the integration of protein sequences and drug chemical structures, Molecules 22 (7) (2017) 1119.
[35] K.M. Giacomini, et al., When good drugs go bad, Nature 446 (7139) (2007) 975.
[36] K. Roy, S. Kar, R.N. Das, Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment, Academic press, 2015.
[37] G. Jin, S.T. Wong, Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines, Drug Discov. Today 19 (5) (2014) 637–644.
[38] J.-F. Wang, D.-Q. Wei, K.-C. Chou, Pharmacogenomics and personalized use of drugs, Curr. Top. Med. Chem. 8 (18) (2008) 1573–1579.
[39] D.-Q. Wei, et al., Molecular modeling of two CYP2C19 SNPs and its implications for personalized drug design, Protein Pept. Lett. 15 (1) (2008) 27–32.
[40] M.A. Johnson, G.M. Maggiora, Concepts and Applications of Molecular Similarity, Wiley, 1990.
[41] D. Butina, M.D. Segall, K. Frankcombe, Predicting ADME properties in silico: methods and models, Drug Discov. Today 7 (11) (2002) S83–S88.
[42] E. Byvatov, et al., Comparison of support vector machine and artificial neural network systems for drug/nondrug classification, J. Chem. Inf. Comput. Sci. 43 (6) (2003) 1882–1889.
[43] H. Li, et al., TarFisDock: a web server for identifying drug targets with docking approach, Nucleic Acids Res. 34 (suppl_2) (2006) W219–W224.
[44] A.C. Cheng, et al., Structure-based maximal affinity model predicts small-molecule druggability, Nat. Biotechnol. 25 (1) (2007) 71.
[45] G. Pujadas, et al., Protein-ligand docking: a review of recent advances and future perspectives, Curr. Pharm. Anal. 4 (1) (2008) 1–19.
[46] M.A. Yıldırım, et al., Drug—target network, Nat. Biotechnol. 25 (10) (2007) 1119.
[47] S.J. Opella, Structure determination of membrane proteins by nuclear magnetic resonance spectroscopy, Annu. Rev. Anal. Chem. 6 (2013) 305–328.
[48] Z. Mousavian, A. Masoudi-Nejad, Drug–target interaction prediction via chemogenomic space: learning-based methods, Expert Opin. Drug Metab. Toxicol. 10 (9) (2014) 1273–1287.
[49] V. Law, et al., DrugBank 4.0: shedding new light on drug metabolism, Nucleic

Acids Res. 42 (D1) (2013) D1091–D1097.
[50] M. Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 28 (1) (2000) 27–30.
[51] E.E. Bolton, et al., PubChem: integrated platform of small molecules and biological activities, Annual Reports in Computational Chemistry, Elsevier, 2008, pp. 217–241.
[52] U. Consortium, UniProt: a hub for protein information, Nucleic Acids Res. 43 (D1) (2014) D204–D212.
[53] R.D. Finn, et al., Pfam: the protein families database, Nucleic Acids Res. 42 (D1) (2013) D222–D230.
[54] N. Hecker, et al., SuperTarget goes quantitative: update on drug–target interactions, Nucleic Acids Res. 40 (D1) (2011) D1113–D1117.
[55] S. Günther, et al., SuperTarget and Matador: resources for exploring drug-target relationships, Nucleic Acids Res. 36 (suppl_1) (2007) D919–D922.
[56] Y. Okuno, et al., GLIDA: GPCR—ligand database for chemical genomics drug discovery—database and tools update, Nucleic Acids Res. 36 (suppl_1) (2007) D907–D912.
[57] C. Qin, et al., Therapeutic target database update 2014: a resource for targeted therapeutics, Nucleic Acids Res. 42 (D1) (2013) D1118–D1123.
[58] M. Kuhn, et al., STITCH 4: integration of protein–chemical interactions with user data, Nucleic Acids Res. 42 (D1) (2013) D401–D407.
[59] A. Gaulton, et al., ChEMBL: a large-scale bioactivity database for drug discovery, Nucleic Acids Res. 40 (D1) (2011) D1100–D1107.
[60] M.P. Magariños, et al., TDR Targets: a chemogenomics resource for neglected diseases, Nucleic Acids Res. 40 (D1) (2011) D1118–D1127.
[61] Z. Gao, et al., PDTD: a web-accessible protein database for drug target identification, BMC Bioinf. 9 (1) (2008) 104.
[62] M. Kuhn, et al., A side effect resource to capture phenotypic effects of drugs, Mol. Syst. Biol. 6 (1) (2010) 343.
[63] D. Emig, et al., Drug target prediction and repositioning using an integrated network-based approach, PLoS One 8 (4) (2013) e60618.
[64] P. Ausloos, et al., The critical evaluation of a comprehensive mass spectral library, J. Am. Soc. Mass Spectrom. 10 (4) (1999) 287–299.
[65] J.-L. Perret, P. Mahe, J.-P. Vert, Chemcpp: an open source c++ toolbox for kernel functions on chemical compounds, 2007. Software available at http://chemcpp.sourceforge.net.
[66] A. Mauri, et al., Dragon software: an easy approach to molecular descriptor calculations, Match 56 (2) (2006) 237–248.
[67] R. Guha, The CDK Descriptor Calculator, NIH Chemical Genomics Center, Indiana, USA, 1991.
[68] N.M. O'Boyle, et al., Open Babel: an open chemical toolbox, J. Cheminfo. 3 (1) (2011) 33.
[69] G. Landrum, RDKit: Open-source cheminformatics, 2006.
[70] C.W. Yap, PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints, J. Comput. Chem. 32 (7) (2011) 1466–1474.
[71] H. Georg, BlueDesc-molecular Descriptor Calculator, University of Tübingen, Tübingen, 2008.
[72] J. Dong, et al., ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation, J Cheminfo. 7 (1) (2015) 60.
[73] D.-S. Cao, et al., Rcpi: R/Bioconductor package to generate various descriptors of proteins, compounds and their interactions, Bioinformatics 31 (2) (2014) 279–281.
[74] D.-S. Cao, et al., PyDPI: Freely Available Python Package for Chemoinformatics, Bioinformatics, and Chemogenomics Studies, ACS Publications, 2013.
[75] N. Xiao, et al., protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences, Bioinformatics 31 (11) (2015) 1857–1859.
[76] B.A. van den Berg, et al., SPiCE: a web-based tool for sequence-based protein classification and exploration, BMC Bioinf. 15 (1) (2014) 93.
[77] D.-S. Cao, Q.-S. Xu, Y.-Z. Liang, Propy: a tool to generate various modes of Chou's PseAAC, Bioinformatics 29 (7) (2013) 960–962.
[78] Y.B. Ruiz-Blanco, et al., ProtDCal: a program to compute general-purpose-numerical descriptors for sequences and 3D-structures of proteins, BMC Bioinf. 16 (1) (2015) 162.
[79] E. Gasteiger, et al., Protein identification and analysis tools on the ExPASy server, The Proteomics Protocols Handbook, Springer, 2005, pp. 571–607.
[80] J.R. Bock, D.A. Gough, Predicting protein–protein interactions from primary structure, Bioinformatics 17 (5) (2001) 455–460.
[81] S.M. Gomez, W.S. Noble, A. Rzhetsky, Learning to predict protein–protein interactions from protein sequences, Bioinformatics 19 (15) (2003) 1875–1881.
[82] S. Martin, D. Roe, J.-L. Faulon, Predicting protein–protein interactions using signature products, Bioinformatics 21 (2) (2004) 218–226.
[83] V.V. Zernov, et al., Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions, J. Chem. Inf. Comput. Sci. 43 (6) (2003) 2048–2056.
[84] S.J. Swamidass, et al., Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity, Bioinformatics 21 (suppl_1) (2005) i359–i368.
[85] R. Guha, A. Bender, Computational Approaches in Cheminformatics and Bioinformatics, John Wiley & Sons, 2011.
[86] M.S. Venkatarajan, W. Braun, New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical–chemical properties, Mol. Model. Ann. 7 (12) (2001) 445–453.
[87] M. Svensén, C.M. Bishop, Robust Bayesian mixture modelling, Neurocomputing 64 (2005) 235–252.
[88] J.-L. Faulon, Stochastic generator of chemical structure. 1. Application to the structure elucidation of large molecules, J. Chem. Inf. Comput. Sci. 34 (5) (1994) 1204–1218.
[89] A. Bender, et al., Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance, J. Chem. Inf. Comput. Sci. 44 (5) (2004) 1708–1718.
[90] J.-L. Faulon, M.J. Collins, R.D. Carr, The signature molecular descriptor. 4. Canonizing molecules using extended valence sequences, J. Chem. Inf. Comput. Sci. 44 (2) (2004) 427–436.
[91] B. Schölkopf, A.J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT press, 2002.
[92] L. Ralaivola, et al., Graph kernels for chemical informatics, Neural Networ. 18 (8) (2005) 1093–1110.
[93] K.M. Borgwardt, et al., Protein function prediction via graph kernels, Bioinformatics 21 (suppl_1) (2005) i47–i56.
[94] T. Evgeniou, C.A. Micchelli, M. Pontil, Learning multiple tasks with kernel methods, J. Mach. Learning Res. 6 (Apr) (2005) 615–637.
[95] E.C. Webb, Enzyme Nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes, Academic Press, 1992.
[96] M.A. Wiering, L.R. Schomaker, Multi-layer Support Vector Machines. Regularization, Optimization, Kernels, and Support Vector Machines, 2014, p. 457.
[97] Y. Xue, et al., Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents, J. Chem. Inf. Comput. Sci. 44 (5) (2004) 1630–1638.
[98] C. Wang, et al., PSoL: a positive sample only learning algorithm for finding non-coding RNA genes, Bioinformatics 22 (21) (2006) 2590–2596.
[99] Y.-W. Chang, et al., Training and testing low-degree polynomial data mappings via linear SVM, J. Mach. Learning Res. 11 (Apr) (2010) 1471–1490.
[100] C.S. Leslie, et al., Mismatch string kernels for discriminative protein classification, Bioinformatics 20 (4) (2004) 467–476.
[101] M. Kumar, V. Thakur, G.P. Raghava, COPid: composition based protein identification, In Silico Biol. 8 (2) (2008) 121–128.
[102] Z.-R. Li, et al., PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence, Nucleic Acids Res. 34 (suppl_2) (2006) W32–W37.
[103] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, J. Mach. Learning Res. 13 (Feb) (2012) 281–305.
[104] D.R. Cox, The regression analysis of binary sequences, J. Roy. Stat. Soc. Series B (Methodological) (1958) 215–242.
[105] J. Zhu, et al., 1-norm support vector machines, Adv. Neural Info. Process. Syst. (2004).
[106] A.Z. Broder, et al., Min-wise independent permutations, J. Comput. Syst. Sci. 60 (3) (2000) 630–659.
[107] S. Niwattanakul, et al., Using of Jaccard coefficient for keywords similarity, Proceedings of the International MultiConference of Engineers and Computer Scientists, (2013).
[108] T. Evgeniou, M. Pontil, T. Poggio, Regularization networks and support vector machines, Adv. Comput. Math. 13 (1) (2000) 1.
[109] T.G. Dietterich, Ensemble methods in machine learning, International Workshop on Multiple Classifier Systems, Springer, 2000.
[110] M. Pal, Random forest classifier for remote sensing classification, Int. J. Remote Sens. 26 (1) (2005) 217–222.
[111] L. Breiman, Random forests, Mach. Learning 45 (1) (2001) 5–32.
[112] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 9 (2008) 1263–1284.
[113] G.M. Weiss, Mining with rarity: a unifying framework, ACM Sigkdd Explor. Newsletter 6 (1) (2004) 7–19.
[114] Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms, Chapman and Hall/CRC, 2012.
[115] S. De Jong, SIMPLS: an alternative approach to partial least squares regression, Chemometr. Intell. Lab. Syst. 18 (3) (1993) 251–263.
[116] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, Adv. Neural Info. Process. Syst. (2002).
[117] T. van Laarhoven, E. Marchiori, Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile, PloS One 8 (6) (2013) e66952.
[118] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach. Learning 63 (1) (2006) 3–42.
[119] J. Shen, et al., Estimation of ADME properties with substructure pattern recognition, J. Chem. Inf. Model. 50 (6) (2010) 1034–1041.
[120] X. Yu, et al., Predicting subcellular location of apoptosis proteins with pseudo amino acid composition: approach from amino acid substitution matrix and auto covariance transformation, Amino Acids 42 (5) (2012) 1619–1625.
[121] M.-G. Shi, et al., Predicting protein–protein interactions from sequence using correlation coefficient and high-quality interaction dataset, Amino Acids 38 (3) (2010) 891–899.
[122] M. Gribskov, A.D. McLachlan, D. Eisenberg, Profile analysis: detection of distantly related proteins, Proc. Natl. Acad. Sci. 84 (13) (1987) 4355–4358.
[123] K.-C. Chou, Y.-D. Cai, W.-Z. Zhong, Predicting networking couples for metabolic pathways of Arabidopsis, EXCLI J. 5 (2006) 55–65.
[124] K.C. Chou, Prediction of protein cellular attributes using pseudo-amino acid composition, Proteins Struct. Funct. Bioinf. 43 (3) (2001) 246–255.
[125] J.M. Keller, M.R. Gray, J.A. Givens, A fuzzy k-nearest neighbor algorithm, IEEE Trans. Syst. Man Cybernetics 4 (1985) 580–585.
[126] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria

of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.
[127] C.J. Ter Braak, Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis, Ecology 67 (5) (1986) 1167–1179.
[128] D.M. Witten, R. Tibshirani, T. Hastie, A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, Biostatistics 10 (3) (2009) 515–534.
[129] T.T. Tanimoto, IBM Internal Report, Nov, 1957, 17, p. 1957.
[130] M.J. Keiser, et al., Relating protein pharmacology by ligand chemistry, Nat. Biotechnol. 25 (2) (2007) 197.
[131] P. Jaccard, Nouvelles recherches sur la distribution florale, Bull. Soc. Vaud. Sci. Nat. 44 (1908) 223–270.
[132] J. Lamb, et al., The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease, Science 313 (5795) (2006) 1929–1935.
[133] F. Iorio, R. Tagliaferri, D.d. Bernardo, Identifying network of drug mode of action by gene expression profiling, J. Comput. Biol. 16 (2) (2009) 241–251.
[134] N. Atias, R. Sharan, An algorithmic framework for predicting side-effects of drugs, Annual International Conference on Research in Computational Molecular Biology, Springer, 2010.
[135] A. Skrbo, B. Begović, S. Skrbo, Classification of drugs using the ATC system (Anatomic, Therapeutic, Chemical Classification) and the latest changes, Medicinski Arhiv 58 (1 Suppl 2) (2004) 138–141.
[136] P. Resnik, Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language, J. Artif. Intell. Res. 11 (1999) 95–130.
[137] T.F. Smith, M.S. Waterman, C. Burks, The statistical distribution of nucleic acid similarities, Nucleic Acids Res. 13 (2) (1985) 645–656.
[138] B.-J. Breitkreutz, et al., he BioGRID interaction database: 2008 update, Nucleic Acids Res. 36 (suppl_1) (2007) D637–D640.
[139] R.M. Ewing, et al., Large-scale mapping of human protein–protein interactions by mass spectrometry, Mol. Syst. Biol. 3 (1) (2007) 89.
[140] J.-F. Rual, et al., Towards a proteome-scale map of the human protein–protein interaction network, Nature 437 (7062) (2005) 1173.
[141] U. Stelzl, et al., A human protein-protein interaction network: a resource for annotating the proteome, Cell 122 (6) (2005) 957–968.
[142] I. Xenarios, et al., DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions, Nucleic Acids Res. 30 (1) (2002) 303–305.
[143] K. Ovaska, M. Laakso, S. Hautaniemi, Fast Gene Ontology based clustering for microarray experiments, BioData Mining 1 (1) (2008) 11.
[144] M. Ashburner, et al., Gene Ontology: tool for the unification of biology, Nat. Genet. 25 (1) (2000) 25.
[145] Y. Liu, et al., DCDB: drug combination database, Bioinformatics 26 (4) (2009) 587–588.
[146] K.-C. Chou, Prediction of G-protein-coupled receptor classes, J. Proteome Res. 4 (4) (2005) 1413–1418.
[147] X. Xiao, P. Wang, K.-C. Chou, iNR-PhysChem: a sequence-based predictor for identifying nuclear receptors and their subfamilies via physical-chemical property matrix, PloS One 7 (2) (2012) e30869.
[148] I. Roterman, et al., Two-intermediate model to characterize the structure of fast-folding proteins, J. Theor. Biol. 283 (1) (2011) 60–70.
[149] X. Xiao, P. Wang, K.-C. Chou, GPCR-2L: predicting G protein-coupled receptors and their types by hybridizing two different modes of pseudo amino acid compositions, Mol. BioSyst. 7 (3) (2011) 911–919.
[150] M.E. Tipping, Sparse Bayesian learning and the relevance vector machine, J. Mach. Learning Res. 1 (Jun) (2001) 211–244.
[151] I. Schomburg, et al., BRENDA, the enzyme database: updates and major new developments, Nucleic Acids Res. 32 (suppl_1) (2004) D431–D433.
[152] S.F. Altschul, E.V. Koonin, Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases, Trends Biochem. Sci. 23 (11) (1998) 444–447.
[153] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometr. Intell. Lab. Syst. 2 (1–3) (1987) 37–52.
[154] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, 2000.
[155] D. Hristovski, et al., Using literature-based discovery to identify disease candidate genes, Int. J. Med. Inf. 74 (2–4) (2005) 289–298.
[156] I. Lee, et al., Prioritizing candidate disease genes by network-based boosting of genome-wide association data, Genome Res. 21 (7) (2011) 1109–1121.
[157] P. Maji, E. Shah, S. Paul, RelSim: an integrated method to identify disease genes using gene expression profiles and PPIN based similarity measure, Inf. Sci. 384 (2017) 110–125.
[158] S. Zickenrott, et al., Prediction of disease–gene–drug relationships following a differential network analysis, Cell Death Dis. 7 (1) (2017) e2040.
[159] M.-W. Huang, et al., SVM and SVM ensembles in breast cancer prediction, PloS One 12 (1) (2017) e0161501.